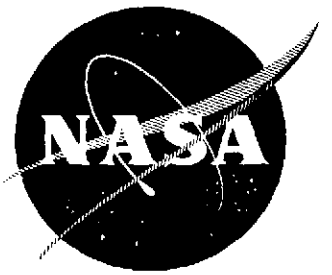
2m14

NASA CR- 134252
ERIM 190100-25-T



IMPROVEMENTS IN ESTIMATING PROPORTIONS OF OBJECTS FROM MULTISPECTRAL DATA

by
H. M. Horwitz
P. D. Hyde
W. Richardson

Infrared and Optics Division

ENVIRONMENTAL
RESEARCH INSTITUTE
OF MICHIGAN
FORMERLY WILLOW RUN LABORATORIES,
THE UNIVERSITY OF MICHIGAN

APRIL 1974

prepared for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
Johnson Space Center, Houston, Texas 77058
Earth Observations Division
Contract NAS 9-9784, Task III

NOTICES

Sponsorship. The work reported herein was conducted by the Environmental Research Institute of Michigan for the National Aeronautics and Space Administration, Johnson Space Center, Houston, Texas 77058, under Contract NAS 9-9784, Task III. Dr. Andrew Potter/TF3 serves as Technical Monitor. Contracts and grants to the Institute for the support of sponsored research are administered through the Office of Contracts Administration.

Disclaimers. This report was prepared as an account of Government-sponsored work. Neither the United States, nor the National Aeronautics and Space Administration (NASA), nor any person acting on behalf of NASA:

(A) Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

(B) Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used above, "person acting on behalf of NASA" includes any employee or contractor of NASA, or employee of such contractor, to the extent that such employee or contractor of NASA or employee of such contractor prepares, disseminates, or provides access to any information pursuant to his employment or contract with NASA, or his employment with such contractor.

Availability Notice. Requests for copies of this report should be referred to:

National Aeronautics and Space Administration
Scientific and Technical Information Facility
P. O. Box 33
College Park, Maryland 20740

Final Disposition. After this document has served its purpose, it may be destroyed. Please do not return it to the Environmental Research Institute of Michigan.

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No. NASA CR-ERIM 190100-25-T	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle IMPROVEMENTS IN ESTIMATING PROPORTIONS OF OBJECTS FROM MULTISPECTRAL DATA		5. Report Date April 1974 6. Performing Organization Code
7. Author(s) H. M. Horwitz, P. D. Hyde, and W. Richardson		8. Performing Organization Report No. ERIM 190100-25-T
9. Performing Organization Name and Address Environmental Research Institute of Michigan Infrared and Optics Division P.O. Box 618 Ann Arbor, MI 48107		10. Work Unit No. Task III 11. Contract or Grant No. NAS 9-9784
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Johnson Space Center Earth Observations Division Houston, Texas 77058		13. Type of Report and Period Covered Technical Report 1 February 1973 through 31 January 1974 14. Sponsoring Agency Code

15. Supplementary Notes

Dr. Andrew Potter/TF3 is Technical Monitor for NASA.

16. Abstract

Methods for estimating proportions of objects and materials imaged within the instantaneous field of view (IFOV) of a multispectral sensor were developed further. These methods are needed when the sensor receives radiation emanating from several objects within the IFOV, and conventional recognition processing hence becomes inadequate.

Improvements in the basic proportion estimation algorithm were devised as well as improved alien object detection procedures. Also, a simplified signature set analysis scheme was introduced for determining the adequacy of signature set geometry for satisfactory proportion estimation. Averaging procedures used in conjunction with the mixtures algorithm were examined theoretically and applied to artificially generated multispectral data. A computationally simpler estimator was considered and found unsatisfactory. Experiments conducted to find a suitable procedure for setting the alien object threshold yielded little definitive result.

Mixtures procedures were used on a limited amount of ERTS data to estimate wheat proportion in selected areas. Results were unsatisfactory, partly because of the ill-conditioned nature of the pure signature set.

17. Key Words	18. Distribution Statement
Multispectral data processing Proportion estimation Alien object detection Mixtures processing Signature set analysis	Initial distribution is listed at the end of this document.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
UNCLASSIFIED	UNCLASSIFIED	78	7.00

PREFACE

This report describes part of a comprehensive and continuing program of research in multispectral remote sensing of the environment from aircraft and satellites. The research is being carried out for NASA's Lyndon B. Johnson Space Center, Houston, Texas, by the Environmental Research Institute of Michigan (formerly the Willow Run Laboratories, a unit of The University of Michigan's Institute of Science and Technology). The basic objective of this program is to develop remote sensing as a practical tool for obtaining extensive environmental information quickly and economically.

In recent times, many new applications of multispectral sensing have come into being. These include agricultural census-taking, detection of diseased plants, urban land studies, measurement of water depth, studies of air and water pollution, and general assessment of land-use patterns. Yet the techniques employed remain limited by the resolution capability of a multispectral scanner. Techniques described in this report may help to overcome this limitation by enabling either examination of the contents of a given scanner resolution cell or, through averaging, faster estimation of the contents of a larger area.

To date, our work on estimation of proportions has included: (1) extension of the signature concept to a mixture of objects; (2) development of a statistical and geometric model for sets and mixtures of signatures; (3) evaluation of computational methods used to estimate proportions of a mixture by maximum likelihood; (4) creation of a computational technique for assessing the expected accuracy of estimation as a function of the signature set; (5) development of techniques to identify alien objects; (6) testing and evaluating the proportion estimation algorithms on artificial as well as actual multispectral scanner data; (7) examining the problem of establishing signatures when pure samples of the objects of interest are not available; and (8) evaluation of alternative estimators.

The research covered in this report was performed under Contract NAS9-9784, Task III, and covers the period from 1 February 1973 through 31 January 1974. Dr. Andrew Potter has been Technical Monitor for NASA. The program was directed by R. R. Legault, Vice-President of the Environmental Research Institute of Michigan (ERIM); J. D. Erickson, Principal Investigator and Head of the ERIM Information Systems and Analysis Department; and R. F. Nalepka, Head of the ERIM Multispectral Analysis Section. The ERIM number for this report is 190100-25-T.

The authors acknowledge the direction provided by Mr. R. R. Legault, Dr. J. D. Erickson, and Mr. R. F. Nalepka; the technical counsel furnished by Mr. R. J. Kauth,

Mr. W. A. Malila, and Dr. R. B. Crane; and the secretarial services of Miss D. Dickerson, Mrs. L. A. Parker, and Mrs. D. Humphrey.

CONTENTS

FIGURES

TABLES

1
SUMMARY

The potential applications of remote sensing appear numerous. However, some of these applications are hampered by the limited spatial resolution of the sensing device. To surmount this difficulty, procedures have been developed for estimating proportions of objects within a single pixel of a multispectral scanner.

This report covers a third phase in the development of proportion estimation techniques. The first two phases included formulation of a solution to the problem, development of suitable algorithms for effective computation, demonstration of feasibility on more or less artificial data, and the pinpointing of problem areas.

This third phase in the development of proportion estimation techniques has focused on problems tending to limit operational usefulness. The speed of the proportion estimation program was increased by theoretical advances in the basic algorithm and by improvements in alien object detection procedures. Other attempts to increase speed included studies of averaging procedures for estimating proportions of objects in a group of pixels and in the use of a simplified proportion estimator. The problem of setting the alien object threshold parameter was studied, but an effective solution not obtained.

In addition, proportion estimation techniques were tested on a limited amount of ERTS-1 data where sufficient ground truth data were available. Results, though disappointing, indicate the need for technique modifications and further experimentation with real data sets.

2
INTRODUCTION

In recent years the staff at ERIM has participated in the development of various techniques for using multispectral remote sensing in many applications, including agricultural land use measurement, rock identification, and water depth measurement.

In conventional multispectral recognition, the total area of each ground material is measured by identifying the material in each ground area (pixel) covered by one resolution element of a multispectral scanner. The total area covered by a ground material is found by adding up the pixels identified with that material. If almost every pixel in the ground scene contains just one of the possible materials, this technique provides adequate estimates of acreages. However, if the pixel contains more than one material in substantial amounts, the pixel cannot be properly classified. For ERTS-1 satellite data, for example, in which each pixel covers about 1.1 acres, the number of pixels containing more than one material may approach 30% of the total area for agricultural crops in the corn belt.

The purpose of the present effort is to obtain improved area estimates of ground materials in these cases. We attempt to overcome the problem of boundary pixels by estimating the proportions of materials within each single pixel.

Since its inception, this effort has consisted of a mix of theoretical model studies and tests with both simulated data and modest amounts of ground-truthed real data. Now that real data sets with adequate associated ground truth are becoming available, we will be able to utilize these in testing and development of mixtures procedures. The past history of the effort is summarized below to provide a context for this report.

Our work on estimation of proportions was accomplished in several phases. In the first phase [1, 2], a mathematical model was constructed which related the multispectral signatures of a mixture to the signatures of component materials. This model permitted the maximum likelihood estimate of the proportion vector to be formulated in terms of the observed data point. The computational aspects of the problem required this simplification: that all of the covariance matrices of the signatures of the component materials be taken as equal to their average. Theoretical and empirical results supported the validity of this assumption. With this simplification, proportion estimation becomes a quadratic programming problem. Several existing computational methods of quadratic programming were adapted and tested on simulated scanner data. Results indicated that this method for proportion estimation was feasible.

The second phase of the program [3] included investigating the problem of detecting alien objects—i.e., objects in the scene not represented in the signature set. A procedure was devised for rejecting those pixels probably containing significant amounts of alien materials. In addition, aircraft scanner data were smoothed over ERTS-sized resolution elements to

simulate spaceborne scanner data. When proportion estimation techniques were tested on this data, estimates of crop acreage based on the estimated proportions were found to be better than estimates obtained with conventional recognition techniques.

The third phase, covering the period of this report, was devoted to investigating problems in proportion estimation that limit its operational usefulness—namely computation time and inaccuracy stemming from the intrusion of alien objects. In addition, some preliminary trials were made with ERTS-1 data. Increased speed was achieved by improvement in the basic algorithm, by conversion to a more advanced computer, and by the further development of data averaging prior to estimating proportions. A reduction in computation time by a factor of about seven was attributed to improvements in the basic algorithm; the advanced computer employed contributed a reduction by a factor of about two. In absolute terms, it required about 20 msec to estimate the proportions of five materials from 12-channel data. Averaging improved the speed of estimation by a factor approximately equal to the number of data points included in the average. (Results of theoretical analyses and simulated tests indicate that averaging may, under certain conditions, even improve the accuracy of proportion estimates.) With the improved proportion estimation algorithm, an ERTS frame containing about 10^7 pixels can be processed in about two hours by averaging signals in groups of 26. It would take over four hours to classify the pixels of the ERTS frame into seven signature categories by recognition processing using a linear decision rule.

Also, in order to achieve increased speed, a theoretically less satisfactory but computationally more tractable estimator of proportions was investigated. Tests on simulated data indicated that although computation time decreased by about 1/3, the estimate's mean-square error increased by about 1/3 on a pixel-by-pixel basis. A surprising result of tests with simulated space data indicated that the new estimate could improve accuracy when estimates are averaged over a sufficient number of pixels (in the case of these tests, more than 27).

In addition, during this phase, a new procedure was devised for detecting alien objects. This procedure depends upon setting the value of a single parameter called the alien object threshold. Tests with simulated data to determine a satisfactory method of setting this threshold were inconclusive.

During the course of the year the mixtures algorithm was applied to three ERTS-1 data sets. Two of these—involving, respectively, lakes and rice fields—are reported elsewhere [4, 5]. The third case, a preliminary investigation of the use of the mixtures algorithm in a general agricultural situation utilizing data gathered under the CITARS (Crop Identification Technology Assessment for Remote Sensing) program, is reported here in Section 7.

The results reported for lakes and for rice fields are most encouraging. These were achieved, however, by employing some special procedures appropriate to the situations. In the

case of lakes, a special detection threshold was set to keep slight indications of water in many pixels from accumulating to a large value. In the case of rice, mixtures were calculated only on those elements located on the boundaries between the rice fields.

In the case of CITARS, the data represent a low contrast scene relative to the other two cases, so no special procedures were used. The results to date on this set of CITARS data are not impressive. It is evident that in order to develop a mixtures algorithm which will be automatic and effective when the spectral information is limited, or during seasons when contrast is low, additional procedures will have to be devised to complement the present algorithm. The only practical way to achieve this will be by utilizing large data sets with excellent ground truth. The past lack of such data has imposed a major constraint on the development of the proportion estimation technique. We are hopeful that the CITARS data will fill this need.

The next section (Section 3) summarizes our approach to the proportion estimation problem, describing improvements in the basic algorithm and some associated computations. Theoretical and numerical results relating to proportion estimation using averaging procedures are presented in Section 4. In Section 5, a simplified estimator of proportions is introduced and compared with the standard estimator. Section 6 reviews our studies of the effect of the alien object threshold setting. Section 7 presents preliminary results of processing a limited amount of ERTS-1 data in an agricultural situation. The appendices contain proofs of theorems and descriptions of computer programs associated with estimation of proportions.

3
APPROACH TO PROPORTION ESTIMATION

This section summarizes ERIM's basic approach to the proportion estimation problem. (See also Refs. [1, 2, 3] for more detail.) The remainder of this section is devoted to the now rewritten computer program (MIXMAP). Improvements made during the reporting period in speed of calculation are introduced in Section 3.2, and the necessary programming is detailed in Appendix A. Modifications of the alien object test to improve both speed and accuracy are described in Section 3.3. (Results of numerical trials on simulated data are reserved for Section 6.) Finally in Section 3.4, under the heading of signature analysis, some preliminaries to the basic proportion estimation calculation are described. Signature analysis provides a measure of the expected quality of the estimates to be obtained from a given signature set.

3.1 MODEL FOR SIGNATURES OF MIXTURES

When the IFOV of a multispectral scanner is large with respect to the structure of the scene being scanned, a single resolution cell or pixel may contain more than a single object or material. A mathematical model has been constructed which relates the signature of a mixture of materials to the signatures of the component materials. Suppose the scanner has n spectral channels and that the signature of object class i, where $1 \leq i \leq m$, is represented by the n-dimensional Gaussian distribution with mean A_i and covariance matrix M_i. Let the proportion of object class i be λ^i and let λ be the vector $(\lambda^1, \lambda^2, \ldots, \lambda^m)^t$, where the superscript t denotes transpose. The signature of the mixture with proportion vector λ is taken to be a Gaussian distribution, with mean A_λ and covariance matrix M_λ given by

$$A_\lambda = \sum \lambda^i A_i = A\lambda$$

$$M_\lambda = \sum \lambda^i M_i$$

where A is the matrix with i column A_i. These formulas constitute our model for signatures of mixtures of materials in terms of signatures of the individual materials.

3.2 ESTIMATION OF PROPORTIONS

The model for a mixture signature can be used to estimate the proportion vector λ corresponding to a signal data vector from a multispectral scanner. Let y denote the n-dimensional data vector from the scanner. A maximum likelihood estimate of the proportion vector λ [2] is a value of λ which minimizes

$$F(\lambda) = \ell n|M_\lambda| + \langle y - A_\lambda, M_\lambda^{-1}(y - A_\lambda) \rangle$$

subject to the constraints that

$$\sum \lambda^i = 1 \text{ and } \lambda^i \geq 0 \quad \text{for} \quad 1 \leq i \leq m$$

Here $|M|$ denotes the determinant of M, M^{-1} is its inverse, and $\langle u, v \rangle$ denotes the inner or dot product of the vectors u and v.*

In general, minimizing $F(\lambda)$ subject to the given constraints is quite difficult. Investigations [2] showed that a good approximation to the minimal λ could be obtained if a simplifying assumption is made. The assumption is that the average of the covariance matrices of the pure signatures can be substituted for each M_i. By using the simplifying assumption and applying a linear transformation which reduces the common covariance matrix to the identity, the problem of estimating λ becomes one of minimizing a function $G(\lambda)$ of the form

$$G(\lambda) = ||y - A_\lambda||^2$$

subject to the constraints on λ. Now y represents the transformed data point, and A_λ the mean of the signature associated with the proportion vector λ after the pure signature means have also been transformed.

The problem of minimizing $G(\lambda)$ subject to the constraints on λ can be viewed geometrically. The set of points $A_\lambda = A\lambda$, where λ is a proportion vector, is the convex hull of the A_i and is called the signature simplex. The problem is to find a proportion vector $\hat{\lambda}$ such that $A\hat{\lambda}$ is the point in the signature simplex closest to the data point y.

The optimal $\hat{\lambda}$ will be unique if the signature simplex is non-degenerate, i.e., has positive m - 1 dimensional volume. This is equivalent to the (n+1)-dimensional vectors $\left(A_i^t, 1\right)$ being linearly independent. We shall always assume this to be the case. Non-degeneracy of the signature simplex implies that the number of materials m in the pure signature set does not exceed the number of spectral channels n by more than one.

Another estimator for λ is considered in Section 5; therefore, to avoid confusion, $\hat{\lambda}$ is called the "standard" estimator.

*Salvato [6] has pointed out the relationship of the ERIM proportion estimation model to the TRW mixtures model [7]. The two models are mathematically equivalent under certain conditions.

The problem of minimizing $G(\lambda)$ can be identified as a quadratic programming problem. In our previous work, a number of extant algorithms were adapted, the most successful based on the method of Theil and van de Panne. Details of the method can be found in Refs. [2, 6]. An improved version of this algorithm was programmed for the IBM 7094 computer during this contract period. The theoretical basis and programming details are given in Appendix A. This version is about 14 times faster than the old version which was programmed for the CDC 1604 computer—a factor of 7 is attributable to improved methodology, and a factor of 2 is attributable to the more advanced computer. It now takes about 20 msec to estimate a mixture of 5 materials with 12 channels of data.

The Theil and van de Panne method adapted to the minimization of $G(\lambda)$ subject to the constraints on λ is described in Ref. [2, (Section I.3)]. It depends upon projections of the data point y onto the linear hulls of certain subsets T of vertices A_i of the transformed signature simplex. The linear hull of a finite set of vectors $v_1, \ldots, v_r$ is the set of all vectors of the form

$$\sum_{j=1}^{r} \eta_j v_j \quad \text{with} \quad \sum_{j=1}^{r} \eta_j = 1$$

When the subset of vertices contains all of the A_i, $1 \leq i \leq m$, then the projection of y is $A\lambda$ where $\begin{bmatrix} \lambda \\ \Delta \end{bmatrix}$ is the solution to

$$\begin{bmatrix} \Gamma_{11} & \cdots & \Gamma_{1m} & 1 \\ & \cdots & & \vdots \\ \Gamma_{m1} & \cdots & \Gamma_{mm} & 1 \\ 1 & 1 \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda^1 \\ \vdots \\ \lambda^m \\ \Delta \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_2 \\ \vdots \\ g_m \\ 1 \end{bmatrix} \tag{1}$$

and

$$\Gamma_{ij} = \langle A_i, A_j \rangle, \quad g_i = \langle A_i, y \rangle$$

When the subset T does not contain all the vertices, the projection of y onto the linear hull of the vertices in T is obtained as follows. For each i for which A_i is not in T, delete the i-th row and column of the matrix in the system of equations (1). Also delete λ^i and g_i from the vectors $\begin{bmatrix} \lambda \\ \Delta \end{bmatrix}$ and $\begin{bmatrix} g \\ 1 \end{bmatrix}$ and consider the remaining system. For example, if T contains all the A_i but A_1 and A_3, the linear system corresponding to T becomes:

$$\begin{bmatrix} \Gamma_{22} & \Gamma_{24} & \Gamma_{25} \cdots \Gamma_{2m} & 1 \\ \Gamma_{42} & \Gamma_{44} & \Gamma_{45} \cdots \Gamma_{4m} & 1 \\ \Gamma_{52} & \Gamma_{54} & \Gamma_{55} \cdots \Gamma_{5m} & 1 \\ \vdots & \vdots & \vdots & \vdots \\ \Gamma_{m2} & \Gamma_{m4} & \Gamma_{m5} \cdots \Gamma_{mm} & 1 \\ 1 & 1 & 1 \cdots 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda^2 \\ \lambda^4 \\ \lambda^5 \\ \vdots \\ \lambda^m \\ \Delta \end{bmatrix} = \begin{bmatrix} g_2 \\ g_2 \\ g_5 \\ \vdots \\ g_m \\ 1 \end{bmatrix}$$

The projection y onto the linear hull of the vertices in T is then given by $A\lambda$ where $\lambda^i = 0$ for A_i not in T, and λ^i is the value in the solution of the linear system corresponding to T for A_i in T.

The new program achieves greatly increased speed by precomputing and storing the inverse of the coefficient matrix in the linear system corresponding to each possible subset T. Additional efficiency is attained by utilizing recursive relationships between successive projections required by the Theil and van de Panne method.

3.3 DETECTION OF ALIEN OBJECTS

Estimating proportions of unresolved objects from a signal y is based on the assumption that the signal comes from a pixel which contains a mixture of materials. These materials are represented by known signatures that constitute the pure signature set. If the pixel should contain a material not represented in the signature set, significant additional error in the estimate of proportions may result. The amount of this error depends upon the proportion of these alien materials and the geometric relationship of their signatures to those in the pure signature set. Those materials occurring in a scene but not represented in the pure signature set are referred to as alien materials or alien objects. Procedures have been designed to reduce the error resulting from the presence of alien objects. These procedures take the form of thresholding tests—hence the designation "alien object threshold."

One might attempt to avoid the alien object problem by obtaining signatures for all materials present in the scene. This approach is usually impractical because of the large number of materials present and the impossibility of obtaining definitive signatures for many of them. An alternative is to use essentially a chi-square test as in conventional recognition processing. Some modifications are necessary when averaging procedures are also employed.

The new mixtures program contains improved procedures for dealing with alien objects. These procedures can be described most easily in terms of the pure signature set and signals after a linear transformation has been employed. After this transformation, we assume that

the i-th material in the pure signature set has mean A_i, and its covariance matrix is the identity. Now given a signal (data point) y from a pixel with unknown proportions of various materials, the estimate $\hat{\lambda}$ of the proportion is obtained as follows. Let Z denote the point in the signature simplex closest to y. Then Z may be represented in the form

$$Z = A\hat{\lambda}$$

where $\hat{\lambda}$ is a proportion vector and is taken as the estimate of proportions in the pixel represented by the signal y. In order to apply an alien object test, we ask, "What is the probability that we would have observed the signal with value exceeding y if the true proportion of the pixel was $\hat{\lambda}$?" Assuming Gaussian signature distributions, this amounts to a chi-square test with n degrees of freedom, where n is the number of spectral channels used. The level of significance is determined by a value χ_o^2, which is the alien object threshold. If

$$||y - Z||^2 = ||y - A\hat{\lambda}||^2 > \chi_o^2$$

then the estimate fails the chi-square test; we then say that the pixel contains significant amounts of alien materials and make no estimate of proportions for the pixel in question. If the estimate passes the test, we accept it as the estimate of proportions of materials in the pixel in question. No significant theoretical results have been obtained as to where the threshold level χ_o^2 should be set in practice, but results of empirical tests are presented in Section 6. This situation is analogous to the chi-square threshold setting in conventional recognition, where it was found that a very high threshold setting was most useful.

In effect, the alien object test is as described for proportion estimation on a pixel-by-pixel basis. However, to reduce computation time, a screening test is also employed in practice. This screening test avoids actual computation of the proportion estimate for some fraction of the cases where the estimate would fail the alien object test. The screening test is computationally rapid and works as follows. Let Z denote the point in the signature simplex closest to the signal y. Set

$$d^2 = ||y - Z||^2$$

Let L_A denote the linear hull of the signature simplex, i.e., L_A is the set of points of the form $A\eta$ where the sum of the components of η is one. (Note that the components are not necessarily positive.) Let Py denote the orthogonal projection of the point y onto L_A; i.e., Py is the point in L_A closest to y. Set

$$d_1^2 = ||y - Py||^2$$

$$d_2^2 = ||Py - Z||^2$$

Then

$$d^2 = d_1^2 + d_2^2$$

Now let $Py = A\eta$. If the components of η are non-negative, then set $\delta_2 = 0$. If some component of η is negative, then Py falls outside the signature simplex and there is an extended simplex such that each face is parallel to its corresponding face in the signature simplex, the hyperplane through each face of the extended simplex is at a distance δ_2 from the hyperplane of the corresponding signature simplex, and Py is in a face of this extended simplex. Then

$$\delta_2 \leqq d_d$$

Set

$$\delta^2 = d_1^2 + \delta_2^2$$

Then

$$\delta^2 \leqq d^2$$

If the alien object threshold is χ_o^2, then the screening test

$$\delta^2 > \chi_o^2$$

will eliminate points that would fail the alien object test, but will accept points that may or may not fail it after d^2 is computed. The reason for using the screening test is that δ^2 is easier to compute than d^2, and δ^2 is often a good approximation to d^2.

In certain averaging procedures, the screening test is used exclusively to maintain computational efficiency. It is used as follows. Let $y_1, \ldots, y_N$ be signals from N pixels. Of these N signals, let $y_1', \ldots, y_{N_1}'$ be the signals which pass the alien object screening test. Let $\bar{y}$ be the average of these N_1 signals and let $\hat{\lambda}$ denote the proportion estimate associated with $\bar{y}$. Then $\hat{\lambda}$ is taken as the estimate of proportions of materials in the region represented by aggregate of pixels corresponding to the signals $y_1', \ldots, y_{N_1}'$. The components of $(N_1/N)\,\hat{\lambda}$ are taken as estimates of the proportions of the materials in the region covered by all N pixels. One of the advantages of averaging is that only a single proportion estimate is required for the region represented by many signals. Using d^2 instead of δ^2 for alien object detection in the averaging procedure would require, in effect, estimation of a proportion vector for each individual signal and thus would cause the loss of the speed advantage inherent in averaging.

3.4 SIGNATURE ANALYSIS

The quality of the estimates of proportions one can expect can be determined to a large extent by examining the pure signature set. In recognition processing we know that the quality of results depends upon the distances between pairs of signature means relative to their spreads (covariances). When these distances are large, good results can be expected. Not only is this requirement necessary for good proportion estimates, but a more stringent condition must be satisfied: that no pure signature be close in a probability sense to any signature of a mixture of the other materials.

A feature of the improved proportion estimation program is a simple test called geometric signature analysis. We deal with the transformed signature simplex with vertices A_i, $1 \leq i \leq m$, and assume that the common covariance matrix of all the transformed signatures is the identity. Let r_i be the distance of A_i to the closest point in the face of the signature simplex opposite A_i. The face opposite A_i is the convex hull of all the vertices A_j except for A_i. Then r_i measures the smallest distance, in standard deviation units, of A_i to the mean of a mixture of the other materials in the signature set. If r_i is small, we would expect data points representing A_i to be confused with data points representing mixtures of the other materials. Figure 1 is an example of a signature simplex well-conditioned for proportion estimation. The circles at the vertices indicate the spread of the distributions at the vertices; these circles are formed by points which are one standard deviation away from the vertex. Each vertex is several standard units away from the closest point in the opposite face. Figure 2, on the other hand, is an example of an ill-conditioned signature simplex. The pure signature mean A_1 is less than a standard deviation away from the closest point in the opposite face.

4
DATA AVERAGING

Two techniques designed primarily to increase speed of calculation have been examined theoretically and numerically with simulated data. These are (1) data averaging, reported in this section, and (2) the use of a simplified estimator, reported in Section 5.

Data averaging can be applied to problems which require an estimate of proportions for a larger area—say for a section or quarter-section of land, or for the land belonging to a particular farm. In addition, data averaging may lead, under certain conditions, to a more accurate estimate of proportions over a wide area than does the pixel-by-pixel estimate. However, data averaging has no role when we desire to know the contents of the scene on a pixel-by-pixel basis.

4.1. ANALYSIS

Consider any area over which average proportions of given materials are to be estimated (see Fig. 3). Suppose that there are N resolution elements in the area, and that for each there is a data (signal) vector y_i and a true proportion vector λ_i. We want to estimate the overall proportion vector $\bar{\lambda}$ for the whole region from the data vectors $y_1, y_2, \ldots, y_N$. Then, assuming the pixels are equal in area,

$$\bar{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i$$

Let $\hat{\lambda}$ be the standard estimator corresponding to a signal y. Sometimes, for clarity, we will exhibit the functional dependence by the notation $\hat{\lambda}(y)$. The usual way to estimate $\bar{\lambda}$ is to find $\hat{\lambda}_i$ for each pixel, then compute the estimator $\hat{\bar{\lambda}}$ of $\bar{\lambda}$ by

$$\hat{\bar{\lambda}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\lambda}_i$$

This "point-by-point estimation" of $\bar{\lambda}$ we call Method 1.

An alternative method for estimating $\bar{\lambda}$ is to obtain the average $\bar{y}$ of the signals

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N}y_i$$

and then compute the estimate

$$\hat{\bar{\lambda}} = \hat{\lambda}(\bar{y})$$

This procedure is "estimation with averaging," sometimes referred to as Method 2. For larger values of N, the computation of $\hat{\bar{\lambda}}$ is faster than the computation of $\hat{\bar{\lambda}}$ by a factor of about N,

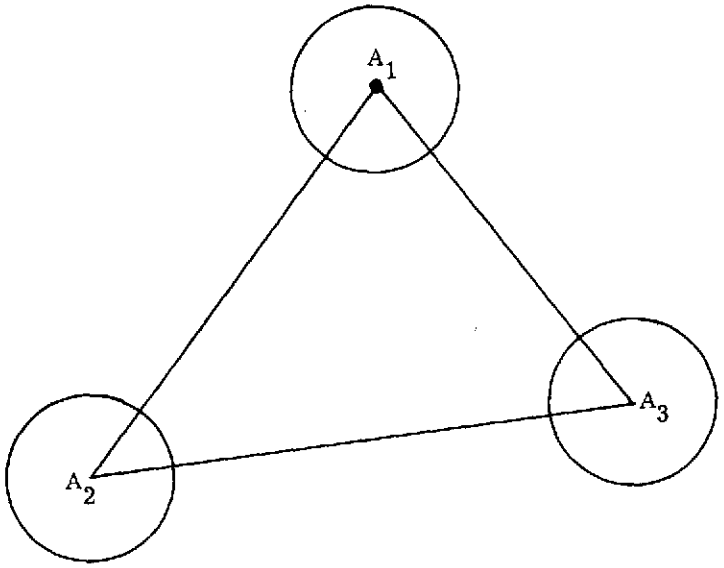


FIGURE 1. WELL-CONDITIONED SIGNATURE SIMPLEX

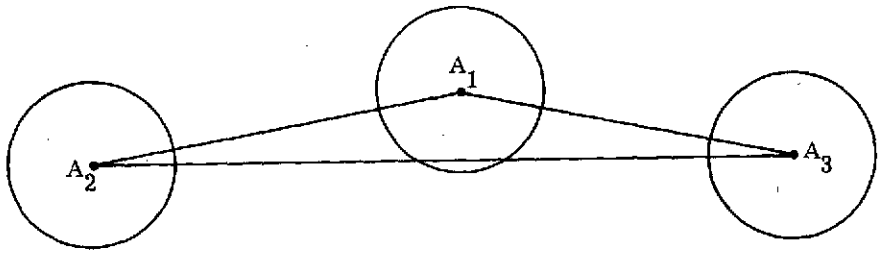


FIGURE 2. ILL-CONDITIONED SIGNATURE SIMPLEX

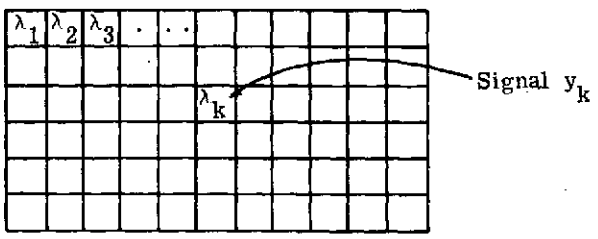


FIGURE 3. SCHEMATIC FOR DATA AVERAGING

since only one proportion estimate is required for Method 2 as opposed to N for Method 1. We can now compare errors in estimates of $\bar{\lambda}$ for the two methods.

The basis of comparison is the mean-square error of the estimate given by

$$\epsilon_1^2 = E\,||\hat{\bar{\lambda}} - \bar{\lambda}||^2$$

for Method 1 (point-by-point) estimation, and by

$$\epsilon_2^2 = E\,||\hat{\bar{\lambda}} - \bar{\lambda}||^2$$

for Method 2 (estimation with averaging) where E is the expected value operator and $||\cdot||$ represents the Euclidean norm. For each i, $1 \leq i \leq N$ let

$$b_i = E(\hat{\lambda}_i) - \lambda_i$$

Then b_i is the bias of the estimate $\hat{\lambda}_i$. Let

$$\bar{b} = \frac{1}{N}\sum_{1}^{N} b_i$$

Then $\bar{b}$ is the average bias of the $\hat{\lambda}_i$. Assuming the y_i are statistically independent, the following result is derived in Appendix B:

$$E\,||\bar{\hat{\lambda}} - \bar{\lambda}||^2 = \frac{1}{N^2}\sum_{1}^{N} E\,||\hat{\lambda}_i - E(\hat{\lambda}_i)||^2 + ||\bar{b}||^2 \tag{2}$$

It then follows that

$$E\,||\bar{\hat{\lambda}} - \bar{\lambda}||^2 \geq ||\bar{b}||^2 \tag{3}$$

Now $\hat{\lambda}_i$ and $E(\hat{\lambda}_i)$ are both proportion vectors; therefore

$$||\hat{\lambda}_i - E(\hat{\lambda}_i)||^2 \leq 2$$

and it follows from Eq. (2) that

$$\epsilon_1^2 = E\,||\hat{\lambda} - E(\hat{\lambda})||^2 \leq \frac{2}{N} + ||\bar{b}||^2 \tag{4}$$

By (3) and (4), the expected mean-square error of the average of the estimates goes to zero as N goes to infinity if and only if the average bias of $\hat{\lambda}_i$ goes to zero; and $\bar{b}$ may not go to 0 as N increases.

On the other hand, it is also shown in Appendix B that

$$\epsilon_2^2 = E\,||\hat{\bar{\lambda}} - \bar{\lambda}||^2 \leq \frac{\beta\tau}{N}$$

where

$$\tau = \max_{i} \text{trace } M_i$$

$$1 \leq i \leq m$$

and β depends upon A but not on N. Note that this result does not depend upon the equal covariance assumption. Thus the expected mean-square error of the proportion estimate for the average signal goes to zero as N goes to infinity.

Since it is reasonable to expect that the average bias $\bar{b}$ will normally remain above a fixed level for typical pure signature sets, $\hat{\bar{\lambda}}$ would be a better estimator of $\bar{\lambda}$ than $\bar{\hat{\lambda}}$ for large N. How large N should be depends on the pure signature configuration. Some results for this will be given below.

4.2 NUMERICAL RESULTS

Using simulated multispectral data based on five agricultural crop signatures, numerical tests were conducted to compare the accuracies of $\bar{\hat{\lambda}}$ and $\hat{\bar{\lambda}}$. The test set contained about 2000 data points; test areas were of size N = 1, 27, 135, 270, 540. Details of the simulation are described in Appendix C.

Estimates of ϵ_1^2 and ϵ_2^2 were obtained, with results as a function of N presented in Table 1. Graphs of these results are presented in Fig. 4. Note that Methods 1 and 2 yield identical estimates for N = 1. Note also that for smaller value of N > 1, Method 1 gives more accurate results. However, in Fig. 4, the two curves cross at about N=330 and Method 2 remains more accurate thereafter. This agrees with the theory of Section 4.1. From the graph for Method 1 we can estimate that the norm-square of the average bias of the estimates $\hat{\lambda}_i$ is about 0.007, the apparent limiting value.

4.3 DISCUSSION AND CONCLUSIONS

The results of the analysis and of the numerical tests, taken at their face value, are in agreement with each other and lead to the same conclusion—namely that for N greater than some threshold value, estimation with averaging should give more accurate results than point-by-point estimation. The value of N for which this occurs is dependent on the particular signature set and on the prior distribution of the proportion vectors. However, a word of caution is in order. This result is based on a highly theoretical situation in which a small, well established set of signatures define the statistical distribution of data points assumed to be uncorrelated. In the world of real data we find that there are always alien materials present. The signal values are not necessarily correlated. Finally, if we attempt to average over a large area to overcome these problems, we may include too many materials in the averaged signal; ultimately, no more than n + 1 materials can be included in the signature set, where n is the number of spectral channels.

Therefore, the details of when averaging should be used and, if it is used, the question of how many pixels should be averaged, will have to be worked out by experience with real data.

TABLE 1. MEAN-SQUARE ERROR IN ESTIMATING $\bar{\lambda}$. (N = number of pixels in area represented by $\bar{\lambda}$.)

Method	N=1	N=27	N=135	N=270	N=540
1 $(\hat{\bar{\lambda}})$	0.2465	0.0145	0.0083	0.0074	0.0073
2 $(\hat{\bar{\lambda}})$	0.2465	0.0385	0.0132	0.0075	0.0052

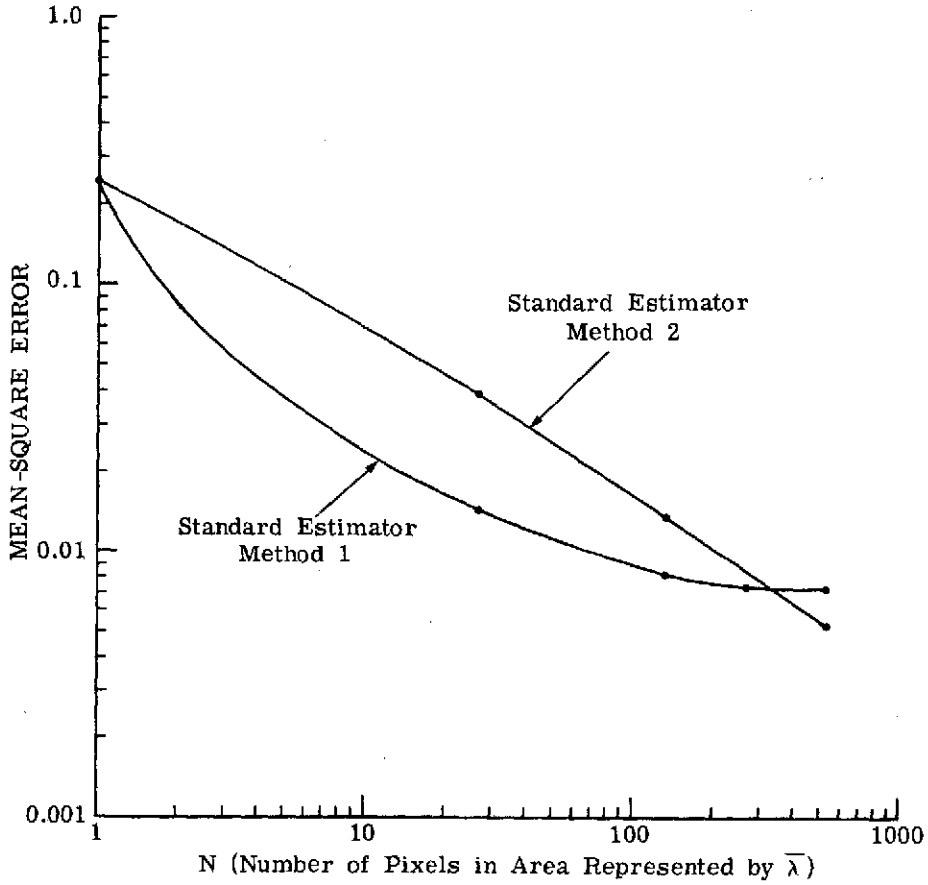


FIGURE 4. MEAN-SQUARE ERROR IN ESTIMATING $\overline{\lambda}$

5
A SIMPLIFIED PROPORTION ESTIMATOR

In another attempt to improve the speed of proportion estimation, a simplified estimator was studied. This estimator is described in Section 5.1 and compared with the standard estimator. The simplified estimator is then analyzed with respect to averaging properties. In Section 5.2, numerical comparisons of the two estimators, based on simulated data, are given with respect to speed of computation and accuracy.

5.1 DESCRIPTION AND ANALYSIS

In Section 3 the standard proportion estimator was defined as the minimum of

$$G(\lambda) = ||y - A_\lambda||^2$$

over λ, subject to the constraints

$$\Sigma\lambda^i = 1 \tag{5}$$

$$\lambda^i \geq 0 \quad \text{and } i = 1, \ldots, m \tag{6}$$

where $\lambda = (\lambda^1, \lambda^2, \ldots, \lambda^m)t$

In this manner the estimate $\hat{\lambda}$ of λ is obtained via quadratic programming. In the simplified estimation procedure the minimization problem is solved subject to constraint (5) but not constraints (6). Solving the problem without constraints (6) is equivalent to finding the projection of y onto the linear hull of the signature mean vectors A_i (vertices of the signature simplex). If this projection falls within the simplex, then constraints (6) are satisfied and the solution λ is the same with or without constraints (6). If this projection falls outside the signature simplex, then the solution will contain negative components. A proportion is then obtained by setting these negative components equal to zero and normalizing the remaining components.

In order to define precisely the simplified estimator $\tilde{\lambda}$, the following definitions are necessary. Let S denote the set of all proportion vectors of dimension n. Let S_A denote the pure signature simplex. Then S_A is the set of vectors of the form $A\lambda$ where λ is in S. Let L be the linear hull of S. Then L is the set of all vectors of dimension n with component sum equal to one. Clearly, L contains S. Let L_A denote the linear hull of S_A. Then L_A is the set of points of the form $A\eta$ where η is in L. For a point η in L, let $\eta^{\oplus}$ denote the vector obtained by setting all negative components of η equal to zero. Let $\rho = \rho(\eta)$ denote the sum of the positive components of η. Note that $\rho > 1$ because the sum of all the components of η is 1. Now let $\tilde{\eta} = \frac{1}{\rho}\eta^{\oplus}$. Note that $\tilde{\eta}$ is a proportion vector.

Let Py denote the orthogonal projection of the signal y onto L_A, i.e., Py is the unique point in L_A closest to y. Py has a representation

$$Py = A\eta$$

for some η in L. This representation is unique for non-degenerate S_A. If λ_o is the true proportion vector of the pixel represented by the signal y then the estimator $\tilde{\lambda}_o$ of λ_o is defined to be

$$\tilde{\lambda}_o = \tilde{\eta}$$

The computational advantage of the estimator $\tilde{\lambda}$ over $\hat{\lambda}$ is that $\tilde{\lambda}$ may be essentially obtained by a single projection. With regard to accuracy, its desirable properties are obscure. If Py falls within the signature simplex S_A then $\tilde{\lambda} = \hat{\lambda}$. But in this case, the computation of $\hat{\lambda}$ using the ERIM-modified Theil and van de Panne algorithm would be essentially the same as the computation of $\tilde{\lambda}$, so that $\tilde{\lambda}$ is not advantageous. When Py is not in S_A, $\tilde{\lambda}$ and $\hat{\lambda}$ may differ considerably. Figures 5(a) and 5(b) show the relationship between $\tilde{\lambda}$ and $\hat{\lambda}$ for two configurations. In Fig. 5(a), $\tilde{\lambda}$ = (3/4, 1/4, 0) and $\hat{\lambda}$ = (1/2, 1/2, 0). In Figure 5(b), $\tilde{\lambda}$ = (1,0,0) and $\hat{\lambda}$ = (0,1,0), the maximal difference.

Some properties of $\tilde{\lambda}$ with respect to averaging procedures will now be examined. These properties are similar to the corresponding averaging properties of the estimator $\hat{\lambda}$. Let y_i, $1 \leq i \leq N$, denote a signal from a pixel with true proportion vector λ_i. The region represented by the aggregate of the N pixels would then have true proportion vector $\bar{\lambda}$ given by

$$\bar{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i$$

Here we are assuming pixels of equal size. If we estimate $\bar{\lambda}$ by averaging the estimates $\tilde{\lambda}_i$ from the individual pixels, we obtain the pixel-by-pixel simplified estimator (Method 1).

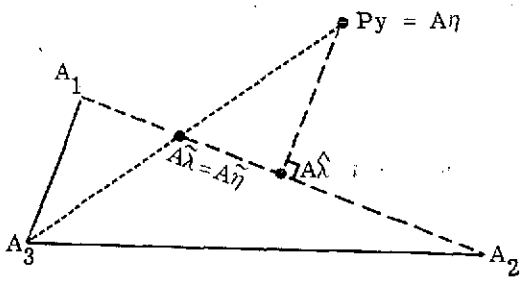
$$\bar{\tilde{\lambda}}_i = \frac{1}{N}\sum_{i=1}^{N}\tilde{\lambda}_i$$

Let

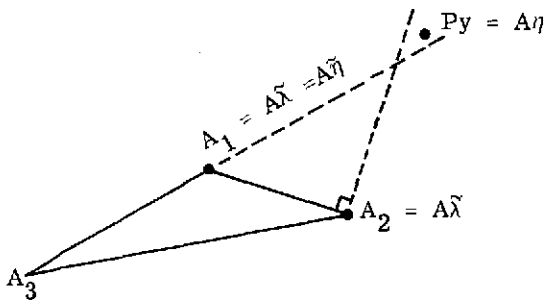
$$\tilde{b}_i = E(\tilde{\lambda}_i) - \lambda_i$$

and let $\bar{\tilde{b}}$ be defined by

$$\bar{\tilde{b}} = \frac{1}{N}\sum_{1}^{N}\tilde{b}_i$$



(a) $\tilde{\lambda} = (3/4, 1/4, 0)$ and $\hat{\lambda} = (1/2, 1/2, 0)$



(b) $\tilde{\lambda} = (1, 0, 0)$ and $\hat{\lambda} = (0, 1, 0)$

FIGURE 5. GEOMETRIC COMPARISON OF STANDARD AND SIMPLIFIED ESTIMATORS

Then $\bar{\tilde{b}}$ is the average bias of the estimators $\tilde{\lambda}_i$, and the mean-square error $\tilde{\epsilon}_1^2$ (assuming as usual that the y_i are statistically independent) is given by

$$\tilde{\epsilon}_1^2 = E\,||\bar{\tilde{\lambda}} - \bar{\lambda}||^2 = \frac{1}{N^2}\sum_1^N ||(\tilde{\lambda}_i - E(\tilde{\lambda}_i)||^2 + ||\bar{\tilde{b}}||^2 \tag{7}$$

It follows that

$$E\,||\bar{\tilde{\lambda}} - \bar{\lambda}|| \geq ||\bar{\tilde{b}}||^2 \tag{8}$$

Since $\tilde{\lambda}_i$ and $E(\tilde{\lambda}_i)$ are both in the simplex S of proportion vectors,

$$||\tilde{\lambda}_i - E\tilde{\lambda}_i||^2 \leq 2$$

and it follows from (7) that

$$\tilde{\epsilon}_1^2 = E\,||\bar{\tilde{\lambda}} - \bar{\lambda}||^2 \leq \frac{2}{N} + ||\bar{\tilde{b}}||^2 \tag{9}$$

By (8) and (9) the expected mean-square error of $\bar{\tilde{\lambda}}$ goes to zero as N goes to infinity if and only if norm-square of the average bias of the estimates $\tilde{\lambda}_i$ goes to zero.

Now let

$$\bar{y} = \frac{1}{N}\sum y_i$$

and let $\tilde{\bar{\lambda}}$ denote the simplified proportion estimate for the signal $\bar{y}$ and consider $\tilde{\bar{\lambda}}$ as an estimate of $\bar{\lambda}$ (Method 2). A proof of the following result is sketched in Appendix B:

$$\tilde{\epsilon}_2^2 = E\,||\tilde{\bar{\lambda}} - \bar{\lambda}|| \leq \frac{m\beta\tau}{N}$$

where m is the number of pure signatures

$$\tau = \max_i \text{trace } M_i$$
$$1 \leq i \leq m$$

and β depends upon A but not N. Thus the expected mean-square error of the proportion estimate $\tilde{\bar{\lambda}}$ for the average signal goes to zero as N goes to infinity.

5.2 NUMERICAL RESULTS

Numerical tests were conducted to compare the standard and simplified estimators. These tests with the simplified estimator corresponded to the tests performed with the standard estimator reported in Section 4.2. Thus areas of size N=1, 27, 135, 270, 540 were used and the simplified estimates of $\bar{\lambda}$ were obtained via Methods 1 and 2. The mean-square error

of the estimates as a function of N are given in Table 2. For N = 1, $\bar{\tilde{\lambda}}$ and $\tilde{\bar{\lambda}}$ coincide, and their entry in the table is the mean-square error for the simplified estimator for a single data point. Comparing this value, 0.3224, with the value 0.2465 for the corresponding entry in Table 1, we can see that for this data, the mean-square error of the simplified estimator is about a third more than the mean-square error for the standard estimator.

Graphs of the results in Fig. 6 show that $\tilde{\epsilon}_2^2$ is going to zero as N increases, as predicted by the theory. It is not clear from the graph what the limiting value $\tilde{\epsilon}_1^2$ is, but this value of $\tilde{\epsilon}_1^2$ would correspond to the norm-square of the average bias $\tilde{b}$ of the estimates $\tilde{\lambda}_1$.

For the purpose of comparing the standard and simplified estimator, the graphs of Figs. 4 and 6 are displayed together in Fig. 7. We see that ϵ_2^2 and $\tilde{\epsilon}_2^2$ eventually coincide. This derives from the fact that for large N, the projection of the average signal $\bar{y}$ falls inside the signature simplex, and this results in the equality of the standard and simplified estimates.

When we compare ϵ_1^2 and $\tilde{\epsilon}_1^2$, we see a surprising result: $\tilde{\epsilon}_1^2$ is actually less than ϵ_1^2 for larger values of N. We cannot explain the phenomenon. We think it may stem from the particular simulated data set employed, and we do not have confidence in the generality of this particular result.

Now let us compare computation time for the simplified and standard estimates. On the basis of estimating proportion for each of the 2000 simulated data points, the simplified estimation procedure required about two-thirds the time required for the standard estimation procedure.

5.3 CONCLUSIONS

The standard estimation is more accurate than the simplified estimator for a single pixel. These two estimators are identical for larger values of N when estimation with averaging (Method 2) is performed. When $\bar{\lambda}$ is estimated by Method 1, we found the simplified estimator superior for larger values of N. The simplified estimator requires about two-thirds the computation time of the standard estimator. We feel that the great drawback of the simplified estimator is the lack of theoretical underpinnings. On balance, we believe that the simplified estimator should be shelved for the present and that proportion estimation development should be continued with the standard estimator.

TABLE 2. MEAN-SQUARE ERROR IN ESTIMATING $\bar{\lambda}$ WITH SIMPLIFIED ESTIMATOR. (N = number of pixels in area represented by $\bar{\lambda}$.)

Method	N=1	N=27	N=135	N=270	N=510
1 $(\tilde{\bar{\lambda}})$	0.3224	0.0128	0.0049	0.0035	0.0027
2 $(\tilde{\bar{\lambda}})$	0.3224	0.0395	0.0137	0.0075	0.0052

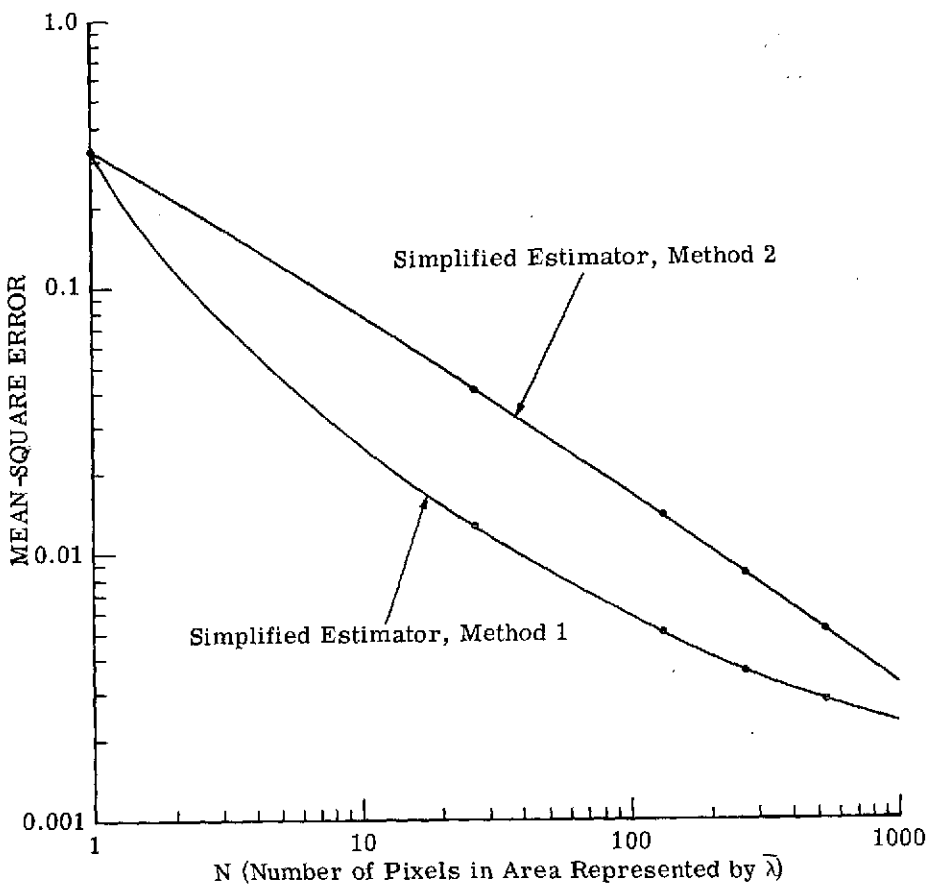


FIGURE 6. MEAN-SQUARE ERROR IN ESTIMATING $\bar{\lambda}$ WITH SIMPLIFIED ESTIMATOR

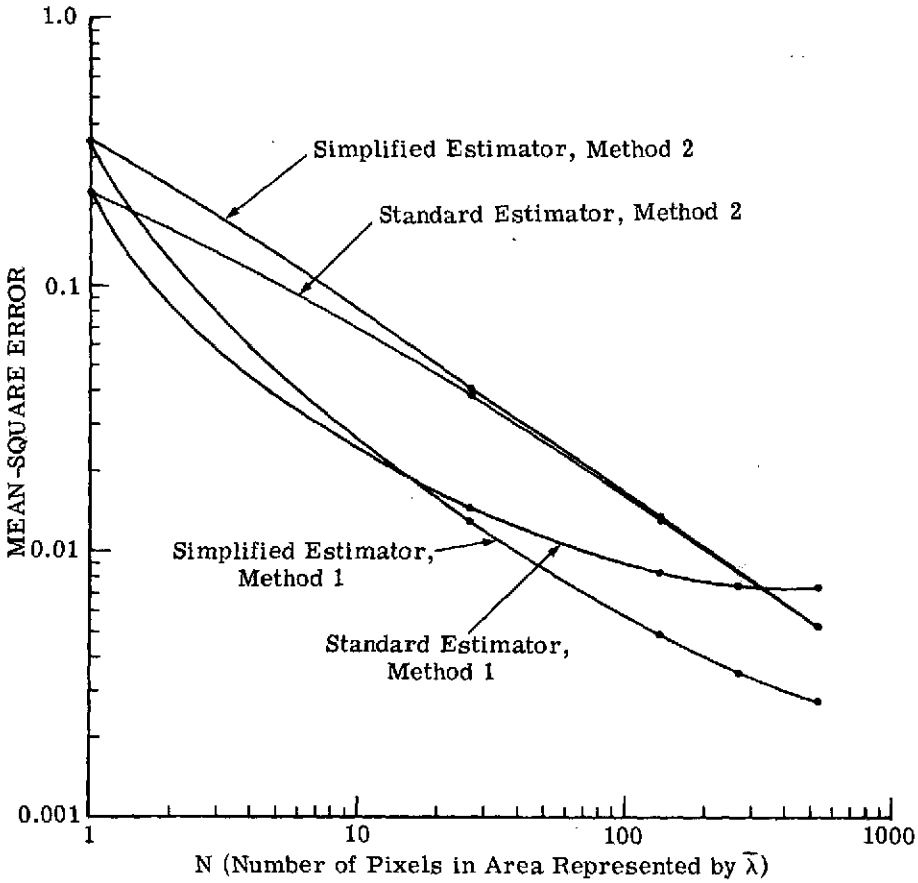


FIGURE 7. COMPARISON OF MEAN-SQUARE ERRORS IN ESTIMATING $\bar{\lambda}$ WITH STANDARD AND SIMPLIFIED ESTIMATORS

6
ALIEN OBJECT THRESHOLD SETTING

In Section 3.3 the improved procedures for dealing with alien objects or materials were described. Alien objects are those objects not represented in the pure signature set. The alien object test employed a generalization of the chi-square test used in recognition processing. In mixtures processing, a signal y is said to represent a pixel with alien material if

$$||y - A\lambda||^2 > \chi_o^2$$

for every proportion vector λ. The alien object threshold is χ_o^2, and in this section we are concerned with determining how best to set the value of this parameter in a specific situation. In Section 3.1, numerical tests were performed to determine the optimal alien object threshold setting.

6.1 NUMERICAL TESTS

The following experiment was performed in order to assess the effect of the alien object threshold setting on estimation accuracy. Three-thousand simulated multispectral data points were generated based on five agricultural signatures: bare soil, corn, soybeans, cut alfalfa, and alfalfa. (The characteristics of the data points are defined in Appendix C.) The data points were then divided into three sets of 1000 points each. Then each set was processed using four of the signatures for the pure signature set, to obtain an estimate of the proportions of these four crops in the region represented by the thousand data points. The point-by-point standard estimate of the average proportion vector for the region represented by the 1000 data points was obtained. This vector had four components; thus the crop not in the signature set played the role of alien material. Then the norm-square of the difference between the region's estimated proportion vector and the true proportion vector was computed and averaged over the three regions represented by the three sets of 1000 data points. The result was taken as an estimate of the mean-square error for the particular alien material and alien object threshold setting chosen. The threshold setting was varied between the values 7 and 37. Graphs of the results are presented in Fig. 8. The graph corresponding to a particular signature set of four crops is identified by the crop treated as alien. Thus, corn alien means that the classes represented in the signature set and used for proportion estimation were bare soil, soybeans, cut alfalfa, and alfalfa.

In Table 3 we list the optimal alien object threshold for each pure signature set—or, equivalently, for a particular alien material. We know that 27 is near-optimal for the two cases: (1) corn alien and (2) alfalfa alien, because a negligible number of data points are rejected at this level.

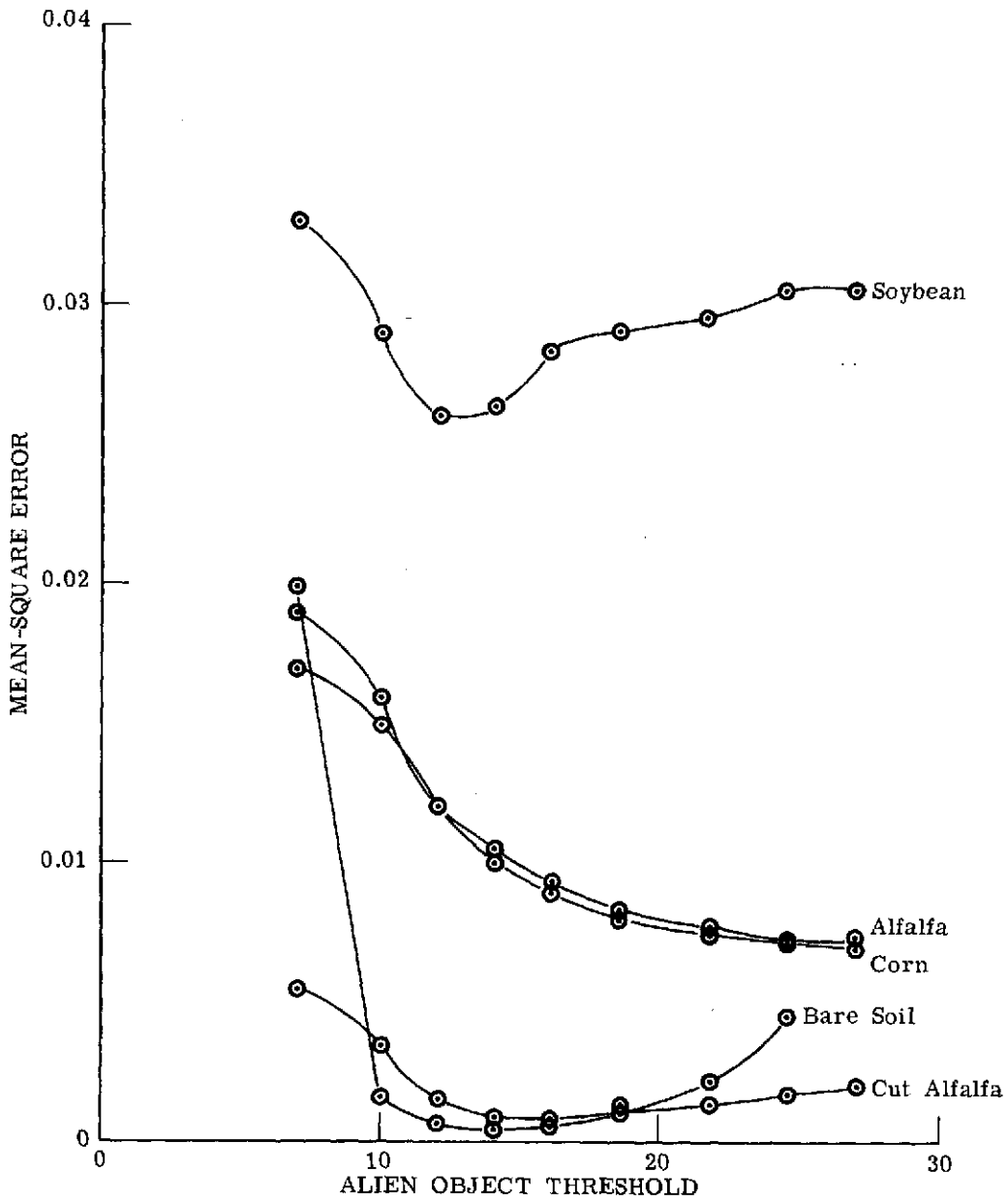


FIGURE 8. ESTIMATED MEAN-SQUARE ERROR OF PROPORTION ESTIMATE AS A FUNCTION OF ALIEN MATERIAL AND ALIEN OBJECT THRESHOLD

The graphs show that a reasonable alien object threshold setting for all five cases would be about 15. We did not find any clear relationship between the distance of the alien signature mean to the corresponding signature set and the optimal threshold value. This may be because the simulated data points were generated with a flat distribution over the set of proportion vectors. As a result the average true proportion vector was about (0.25, 0.25, 0.25, 0.25) in each of the five cases. The geometrical relationship of the alien signature mean to the means of material in the signature set may distort the optimal threshold that would result if several different values of $\bar{\lambda}$ were used. This, we believe, is the reason that two cases in Fig. 8 show no clear-cut finite optimal value.

6.2 CONCLUSION

For the limited tests conducted, we found a single alien-object threshold setting satisfactory in all five cases. This result is based on a very limited experiment. However, the procedure for finding the best threshold setting on simulated data may be used on training sets for real data—in fact, we do this in the next section. (When there are not enough training data available, we still don't have a completely satisfactory way to set the alien object threshold.)

TABLE 3. OPTIMAL ALIEN OBJECT THRESHOLD

Alien Class	Optimal Threshold
Bare Soil	12
Corn	27
Soybeans	12
Cut Alfalfa	16
Alfalfa	27

7
ESTIMATING WHEAT PROPORTIONS FROM ERTS DATA

An exercise was conducted using the mixtures algorithm to determine wheat proportions in selected regions from ERTS-1 multispectral scanner (MSS) data. The data were collected over Fayette County, Illinois, on 11 June 1973. Ground truth information was collected as part of the CITARS (Crop Identification Technology Assessment for Remote Sensing) task, a joint investigation by the Earth Observations Division (EOD) of the Johnson Space Center, Purdue University's Laboratory for Applications of Remote Sensing (LARS), and ERIM.

Specifically, the objective of the exercise was to estimate by mixtures processing the proportion of wheat in each of 20 quarter-sections. Results were disappointing in that the rms error of the estimates was almost 85% of the known proportion of wheat averaged over these quarter-sections. By comparison, the rms error reported for estimates obtained by conventional recognition processing for the same quarter-sections was about 50% [9].

Tests on the signature set used in mixtures processing showed that the signatures of the individual materials were very close in a probability sense to the signatures of mixtures of the other materials. Thus, we should not have expected to obtain very good results. The ill-conditioned nature of the signature set derives to some extent from the limited number (4) of the ERTS-1 sensor's spectral bands, and their relatively large width. Also, the time of year contributed to the low contrast. Nevertheless, results of the exercise indicate that present proportion estimation procedures should be modified for these low-contrast situations.

7.1 DATA FLOW AND SIGNATURE EXTRACTION

Multispectral data flow during processing as well as preliminary data handling steps have been already described [9]. Procedures for extracting signature statistics from training data are given in the same document. We will now discuss the quality of signatures obtained and the procedure for selecting the signature set used in estimating proportions.

Signatures for seven object classes were obtained: wheat, corn, soybeans, clover, trees, pasture, and water. The number of training data points used for estimating signature parameters for each of these crop types is shown in Table 4. Note that only 9 data points were available for estimating the pasture signature and that 17 points were used in estimating the clover signature. The last column of the table lists the determinant of the estimated covariance matrix for each object class. [The square root of the determinant is proportional to the volume of the unit contour ellipse of the class signature and is a measure of the spread of that signature. Considerable variation among the estimated signature covariance matrices may thus be expected.]

A training set consisting of 20 quarter-sections was selected in the following fashion: All of the quarter-sections are in a land segment 5 miles wide by 20 miles long. The 100 sections

comprising this segment were numbered starting with the top left section (No. 1) and proceeding to the top right (No. 5); the next row starts at 6 on the left and goes to 10 on the right, and so on. With each of the 20 quarter-sections carrying the same number as the full section to which it belonged, the 10 odd-numbered quarter-sections in this ordering were then designated as belonging to the training set.

Since the mixtures algorithm permits, at most, five signatures in the signature set when four channels of spectral data are available, two signatures were eliminated by the following method. The distance r of each object class from wheat was computed by geometric signature analysis and the amount, α, of each of the classes in the 10 training quarter-sections was determined. Then the two object types for which

$$\alpha\phi(-r/2)$$

were least were excluded. Here $\phi(x)$ is the univariate Gaussian density function. On this basis the five signatures retained in the signature set represented wheat, corn, soybeans, clover, and pasture.

7.2 SIGNATURE SET ANALYSIS

Analysis of the signature set retained for mixtures processing showed that the set was ill-conditioned—that is, some or all of the pure signature means were close, in a probability sense, to the mean of the signature of some mixture of the other materials. In fact, for this particular case, all of the pure signature means were close to the mean of the signature of some mixture of the other four materials. Table 5 quantifies this separation in standard deviation units for each of the materials in the signature set. Our past experience with the mixtures algorithm indicates that, for satisfactory results, all the distances determined by signature set analysis should be about 3 or more.

7.3 SETTING THE ALIEN OBJECT THRESHOLD

The procedure for setting the alien object threshold was to take nine specified values and then choose that value found to be best for the training data. The nine values initially taken were all too large, so much lower values were then tried; the best alien object threshold value for the training data was found to be 2. This value was used for processing all 20 training quarter-sections.

Only five of the ten training quarter-sections were employed for setting the alien object threshold—namely, those numbered 25, 32, 44, 83, 86. The training quarter-sections numbered 7, 19, 72, 73 were excluded because they contained less than 10% wheat, which we felt would bias the threshold setting to the low side. Training quarter-section No. 56 was excluded because the ground truth at hand was of questionable validity.

TABLE 4. NUMBER OF DATA POINTS USED IN ESTIMATING SIGNATURES AND DETERMINANTS OF ESTIMATED COVARIANCE MATRICES

Object Class	Number of Data Points	Determinant of Estimated Signature Covariance Matrix
Wheat	47	265.5
Corn	88	12934.3
Soybeans	116	750.4
Clover	17	47.4
Trees	180	3.7
Pasture	9	3.6
Water	120	6.5

TABLE 5. RESULTS OF SIGNATURE SET ANALYSIS. (Distance is measured from mean of the material signature to the mean of the mixture signature which is closest to it in a probability sense.)

Material	Distance (Standard Deviation Units)
Wheat	0.22
Corn	0.31
Soybeans	0.24
Clover	0.41
Pasture	0.17

7.4 MIXTURES PROCESSING RESULTS

Table 6 gives the results obtained in using mixtures processing to estimate the proportion of wheat for each of the 20 quarter-sections. For comparative purposes, results of conventional ERIM recognition processing [9] are included. With mixtures processing, the rms error for the estimated wheat percent was about 0.85 of the average wheat percent over the 20 quarter-sections—not good. Since the corresponding rms error was about 0.6 for recognition processing, it is clear that recognition processing outperformed mixtures processing.

7.5 CONCLUSIONS

The fact that mixtures processing gave poorer results than recognition processing in this exercise indicates that mixtures processing techniques require further development for low-contrast situations.

The ill-conditioned nature of the signature set, as revealed by signature set analysis, indicates that the spectral information gathered during certain times of the year by the four ERTS-1 channels is insufficient to disperse the signatures of materials enough to permit successful mixtures processing without modifying present procedures.

TABLE 6. RESULTS OF ESTIMATING WHEAT PROPORTIONS IN 20 QUARTER-SECTIONS BY MIXTURES PROCESSING AND CONVENTIONAL RECOGNITION PROCESSING

Quarter-Section (ERIM No.)	Ground Truth Wheat %	Estimated % Wheat Mixtures Processing	Estimated % Wheat Recognition Processing
7	0	20.3	13.0
18	8.3	11.1	14.0
19	4.4	4.1	2.1
21	28.5	12.5	11.8
25	21.9	19.8	12.4
28	22.9	16.2	21.7
32	15.5	8.0	8.5
40	2.8	11.9	3.0
44	18.4	21.3	18.8
46	0	14.2	17.3
56	26.6	10.0	10.7
64	20.5	4.6	10.9
72	9.3	12.2	14.9
73	0	11.5	15.2
75	0	9.1	8.0
82	22.4	15.7	14.0
83	17.6	10.3	15.3
84	22.5	25.0	35.2
86	12.5	23.0	24.2
100	flooded (0)	18.0	--
Average	12.9	13.8	13.9
RMS Error		11.0	7.9

8
CONCLUSIONS AND RECOMMENDATIONS

The multispectral data processing method for estimating proportions of objects and materials appearing within the instantaneous field of view of a sensor system has been improved —especially with regard to the computational speed of the basic algorithm and the associated alien object detection scheme.

Although the simplified proportion estimator did not prove adequate, results obtained in tests of these procedures on more or less artificial data sets indicate their potential usefulness. Averaging schemes to further increase processing speeds appear to be advantageous under certain circumstances.

Success with applications of proportion estimation techniques on ERTS data have been reported [4, 5]. However, during certain times of the year a low-contrast situation may occur. In such situations, we feel some modification of the techniques will improve performance. One such modification would be based on the assumption that any pixel contains a very limited number of materials, say not more than two or three. In addition, the estimated proportion of a material in a pixel should perhaps be required to have some minimum value before it is counted.

We believe that advances in proportion estimation now require experimentation with real data sets accompanied by adequate associated ground truth information.

Appendix A
DETAILS OF COMPUTATIONAL IMPROVEMENTS

The present adaptation of the Method of Theil and van de Panne is faster than the original version by a factor of 14. This progress is attributable to (1) a faster machine, (2) increased computer memory that obviates recomputation of certain quantities, and (3) discovery of projection theorems that permit computational shortcuts. These newer theorems—which have enabled a number of improvements to be made in the estimation algorithm, the alien object test, and geometric signature analysis—are detailed below.

A.1 THEOREMS LEADING TO IMPROVEMENT

The Method of Theil and van de Panne computes the closest point $\hat{y} = A^T\hat{\lambda}(y)$ to y in the signature simplex by computing certain projections of y onto the simplex and its subsimplices. Though it does not need to compute every such projection, the method starts by projecting y onto the hyperplane determined by the signature mean vectors $A_1, \ldots, A_m$; it then projects y onto hyperplanes of succeedingly lower dimensions. The projection y^* of y onto the hyperplane determined by $A_1, \ldots, A_m$ is computed by obtaining $\lambda = (\lambda_1, \ldots, \lambda_m)^\dagger$ from the equation

$$\begin{bmatrix} \Gamma_{11} & \cdots & \Gamma_{1m} & 1 \\ \vdots & & \vdots & \vdots \\ \Gamma_{m1} & \cdots & \Gamma_{mm} & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_m \\ \Delta \end{bmatrix} = \begin{bmatrix} g_1 \\ \vdots \\ g_m \\ 1 \end{bmatrix} \qquad \text{(A-1)}$$

where $\Gamma_{ij} = A_i^T A_j$ and $g_i = A_i^T y$. Then $y^* = A^T\lambda$. To project y onto a lower-dimensional hyperplane determined by some but not all of the A_i, delete the rows and columns corresponding to the A_j not included, and solve the reduced system of equations for the λ_j. If the projection hyperplane does not include A_j, λ_j must be omitted from the equation and set to zero.

All these projections would ordinarily require a great deal of computation, but the theorems given below afford some savings in effort.

Theorem 1. The square of the distance from y to y^* is $y^Ty - \lambda^Tg - \Delta$.

Proof. D^2, the squared distance, $= (y - y^*)^T(y - y^*)$

$$= (y - A^T\lambda)^T(y - A^T\lambda) = y^Ty - 2\lambda^TAy + \lambda^TAA^T\lambda.$$

Let J be a column vector of m 1's. Equation (A.1) implies

$$\Gamma\lambda + \Delta J = g$$

† In this appendix, λ_i denotes the i-th component of the vector λ and the A_i are row vectors forming rows of the matrix A.

In other words,

$$AA^T\lambda = Ay - \Delta J$$

Substituting in the third term of D^2, D^2 becomes

$$y^Ty - 2\lambda^TAy + \lambda^T(Ay - \Delta J) = y^Ty - \lambda^TAy - \Delta\lambda^TJ = y^Ty - \lambda^Tg - \Delta$$

because $\lambda^TJ = 1$. Q.E.D.

This theorem provides a convenient way of computing the "out-of-plane distance" in the alien object test described.

Let λ' be the λ vector that would be obtained by deleting a vertex A_k from the set of vertices considered. The next theorem shows a relationship between λ and λ'.

Let $\Gamma_+ = \begin{bmatrix} \Gamma & J \\ J^T & 0 \end{bmatrix}$ and $C = \Gamma_+^{-1}$

Because Γ_+ is symmetrical, C is symmetrical. The solution to Eq. (A-1) is

$$\begin{bmatrix} \lambda \\ \Delta \end{bmatrix} = C\begin{bmatrix} g \\ 1 \end{bmatrix} \qquad \text{(A-2)}$$

Theorem 2. $\lambda_i' = \lambda_i - \dfrac{C_{ik}}{C_{kk}}\lambda_k$

$$\Delta' = \Delta - \frac{C_{+k}}{C_{kk}}\lambda_k$$

where the subscript + stands for "m + 1".

Proof. Let C' be the result of inverting a Γ_+ that is missing the k-th row and column. It is a matrix theorem that

$$C_{ij}' = C_{ij} - \frac{C_{ik}C_{jk}}{C_{kk}}$$

This lemma can be verified by multiplying row i of C' by column h of Γ_+'. The result is

$$\sum_{j\neq k} C_{ij}'\Gamma_{+jh}' = \sum_j \left(C_{ij} - \frac{C_{ik}C_{jk}}{C_{kk}}\right)\Gamma_{+jh} = \sum_j C_{ij}\Gamma_{+jh} - \frac{C_{ik}}{C_{kk}}\sum_j C_{ik}\Gamma_{+jh}$$

The term for j = k can be included because the factor in parentheses is 0 when j = k.

$\Sigma C_{jk}\Gamma_{+jh} = 0$ because $C_{jk} = C_{kj}$, for $k \neq h$, and C is the inverse of Γ_+. Thus the second term drops out. The first term is 1 or 0 for h equal or not equal to i, proving the lemma.

Define g_{m+1} to be 1. By Eq. (A-2)

$$\lambda_i' = \sum_{\substack{j=1 \\ j\neq k}}^{m+1} C_{ij}' g_j$$

with the k-th term missing. Therefore

$$\lambda_i' = \sum_{j=1}^{m+1} \left(C_{ij} - \frac{C_{ik}C_{jk}}{C_{kk}} \right) g_j$$

with the k-th term included because, as was seen, it is 0. This equation implies

$$\sum_{j=1}^{m+1} C_{ij}g_j - \frac{C_{ik}}{C_{kk}} \sum_{j=1}^{m+1} C_{jk}g_j = \lambda_i - \frac{C_{ik}}{C_{kk}}\lambda_k \quad \text{Q.E.D.}$$

The proof of the theorem for Δ is analogous.

Because C_{ik} and C_{kk} depend on the A_i values but not on y, they are constants that can be precomputed and stored; Theorem 2 provides a quick way to proceed from λ to λ', thus making possible a considerable speedup in the Thiel and van de Panne method.

Theorem 3. The squared distance d^2 from y^* to the k-th face of the simplex is

$$\frac{\lambda_k^2}{C_{kk}} - \Delta$$

By the k-th face of the simplex is meant the plane through all the vertices A_i except A_k.

Proof. By Theorem 1,

$$d^2 = y^{*T}y^* - \lambda'^T g' - \Delta'$$

$y^{*T}y^* = \lambda^T AA^T\lambda$ because $y^* = A^T\lambda$. Let us define $\lambda_k' = 0$ to keep λ' a convenient dimension. $g' = A'y^*$ where A' is A with the k-th row missing. But let us define $g_k' = A_k y^*$ to make g' the same dimension as λ' and give it the convenient definition Ay^*. The value of $\lambda'^T g'$ is unchanged by this maneuvering.

$$g' = Ay^* = AA^T\lambda$$

Thus

$$d^2 = \lambda^T AA^T \lambda - \lambda'^T AA^T \lambda - \Delta' = (\lambda - \lambda')^T AA^T \lambda - \Delta$$

Now $AA^T\lambda = g - \Delta J$ by Eq. (A-1). By Theorem 2,

$$\lambda - \lambda' = \frac{C_{ik}}{C_{kk}}\lambda_k$$

and

$$\Delta' = \Delta - \frac{C_{+k}}{C_{kk}}\lambda_k$$

so

$$d^2 = \frac{\lambda_k}{C_{kk}}(C_{ik})^T(g - \Delta J) + \frac{\lambda_k}{C_{kk}}C_{+k} - \Delta = \frac{\lambda_k}{C_{kk}}\left(\sum_i C_{ik}g_i + C_{+k} - \Delta\sum_i C_{ik}\right) - \Delta$$

$\sum_i C_{ik} = 0$ because it is the k-th row of C times the last column of Γ_+. $\sum_i C_{ik}g_i + C_{+k} = \lambda_k$,

by Eq. (A-2), hence

$$d^2 = \frac{\lambda_k^2}{C_{kk}} - \Delta \quad \text{Q.E.D.}$$

If λ_k is negative, y* is on the outside of the simplex, i.e., the k-th face separates y* from the simplex. If every λ_k is non-negative, y* is in the simplex. We therefore define a lower bound δ of the distance from y* to the simplex as follows.

$$\delta^2 = \min_{\lambda_k<0}\left(\frac{\lambda_k^2}{C_{kk}}\right) - \Delta \quad \text{if some } \lambda_k < 0 = 0 \quad \text{if all } \lambda_k\text{'s} \geq 0$$

δ is a lower bound because the nearest point on the simplex to a point outside it has to be on some face, say the k-th face of the simplex. The distance from y* to that point $\geq$ the distance from y* to that face $\geq \delta$.

Corollary 4. A lower bound for the square of the distance from y to the simplex is $y^Ty - \lambda^Tg - \Delta + \delta^2$.

Proof. Let s be any point on the simplex. Because y-y* is orthogonal to y*-s, the squared distance from y to s is the squared distance from y to y*, which by Theorem 1 is $y^Ty - \lambda^Tg - \Delta$, plus the squared distance from y* to s, which has been shown to have the lower bound δ^2. Q.E.D.

The next theorem shows conditions under which the Δ term in the distance expressions can be ignored.

Theorem 5. When either y or the origin of the space is in the plane through the vertices, $\Delta \approx 0$.

Proof. If y is in the plane of the simplex, $y = \sum_k \lambda_k A_k$

$$g_i = A_i(\Sigma\lambda_k A_k)^T = \sum_k \lambda_k \Gamma_{ik}$$

$$\Delta = C_{++} + \sum_{i=1}^{m} C_{i+} g_i = C_{++} + \sum_{i=1}^{m} C_{i+} \sum_k \lambda_k \Gamma_{ik}$$

Reversing the order of summation,

$$\Delta = C_{++} + \sum_{k=1}^{m} \lambda_k \sum_{i=1}^{m} C_{i+} \Gamma_{ik}$$

C is symmetric, and, $\sum_{i=1}^{m+1} C_{i+} \Gamma_{ik}$ is the inner product of the m+1 row of C with the k-th row of Γ_+. Hence when the last term of the sum is deleted

$$\sum_{i=1}^{m} C_{i+} \Gamma_{ik} = -C_{++}$$

$$\Delta = C_{++} + \sum_{k=1}^{m} \lambda_k (-C_{++}) = 0$$

because the λ's sum to 1, and the first part of the theorem is proved.

To prove the second part of the theorem, we draw a distinction between the "mixture space," which is all points of the form

$$\lambda_1 A_1 + \ldots + \lambda_m A_m$$

where $\lambda_1 + \ldots + \lambda_m = 1$, though not necessarily all positive, and the "vector space" spanned by the vectors $A_1, \ldots, A_m$. The mixture space is what we have previously called "the plane through the vertices."

We will first show that when the origin is in the mixture space, the mixture space and the vector space are identical. Every mixture is a linear combination of the basis vectors and hence is in the vector space. It remains to show that every point in the vector space is a mixture. Let the origin be the mixture $\lambda_1 A_1 + \ldots + \lambda_m A_m$ and let $V = C_1 A_1 + \ldots + C_m A_m$ be an arbitrary point in the vector space. Let d be $C_1 + \ldots + C_m$.

$$V = C_1A_1 + \ldots + C_mA_m + (1 - d)(\lambda_1A_1 + \ldots + \lambda_mA_m)$$

because the second term is 0.

$$V = [C_1 + (1 - d)\lambda_1]A_1 + \ldots + [C_m + (1 - d)\lambda_m]A_m$$

The coordinates now add up to 1 and V is therefore a mixture.

Now let y be an arbitrary point in the whole space. $y = y^* + y'$ where y^* is in the mixture space and y' is orthogonal to it.

$$y = (\Sigma\lambda_kA_k)^T + y'$$

$$g_i = A_iy = A_i\Sigma\lambda_kA_k^T + A_iy' = \Sigma\lambda_k\Gamma_{ik}$$

The second term drops out because y' is orthogonal to every vector in the mixture space. The remainder of the proof is the same as for the first part of the theorem. Q.E.D.

Corollary 6. The distance from a vertex A_k to the opposite face of the simplex is $1/\sqrt{C_{kk}}$.

Proof. By Theorem Γ, the squared distance from A_k to the k-th face is

$$\frac{\lambda_k^2}{C_{kk}} - \Delta$$

λ_k is 1, and by Theorem 5, $\Delta = 0$. Hence the corollary. Q.E.D.

The application and usefulness of these theorems are illustrated in the remainder of this appendix.

A.2 DESCRIPTION OF THE ALGORITHM

The quadratic programming problem derived from maximum likelihood and the mixtures model is solved by an adaptation of the Method of Theil and van de Panne. Since our previous report, certain improvements have been made in the algorithm. For the sake of completeness, our original version [2] of Theil and van de Panne is outlined here. A description is then given for the improvements which have permitted reduction of computation time.

Let {h} denote the set consisting of the integer h, and let S be a subset of the index set {1, a, . . . , m}. Let the A_j be the vertices of the (transformed) signature simplex. Let H^S be the hyperplane formed by those B_j so that $j \notin S$. Let z^S be the projection of z onto H^S. Then there exists a vector p^S so that

$$z^S = \sum_{j=1}^{m} \lambda_j^S A_j$$

$$\lambda_j^S = 0 \quad \text{for } j \in S$$

$$\sum_j \lambda_j^S = 1$$

However, λ^S may not be a proportion vector, since there may be j so that $\lambda_j^S < 0$. The set of such j is called the violation set V(S) of S. Let ϕ denote the empty set. Obviously, if $V(S) = \phi$, then λ^S is a proportion vector.

As noted before, the Theil and van de Panne method finds the closest point in the simplex to z by computing projections of z onto various hyperplanes H^S, in such a manner that not all such projections have to be made. Let $\hat{\lambda}$ be the desired proportion vector. Then there exists $\hat{S}$ such that $\hat{\lambda} = \lambda^{\hat{S}}$, and the following rules (theorems) apply:

(1) If $V(\phi) \neq \phi$, there is some $h \in V(\phi)$ such that $h \in \hat{S}$.

(2) If $V(S) \neq \phi$, then $S \subset \hat{S}$ only if there is at least one $h \in V(\hat{S})$ so that $h \in S$.

(3) For any S such that $V(S) = \phi$, λ^S is the desired proportion vector only if $h \in V(S - \{h\})$ for all $h \in S$

Given these rules, it is possible to define the algorithm in detail.

The Theil and van de Panne method is divided into, at most, m stages. In stage k, the sets S_k all have exactly k elements. The algorithm proceeds as follows:

Stage 0: Compute λ^{ϕ}. If $V(\phi) = \phi$, then λ^{ϕ} is the desired proportion vector and the algorithm terminates. Otherwise, $\hat{S} \neq \phi$. Proceed to Stage 1.

Stage 1: Using rule (1) above, consider all $S_1 = \{h\}$ so that $h \in V(\phi)$. Compute λ^{S_1} and $V(S_1)$ for all such S_1. If for any S_1, $V(S_1) = \phi$, apply rule (3) to determine whether λ^{S_1} is the desired solution. If it is, the algorithm terminates. If not, S_1 is removed from the collection of sets for Stage 1.

Stage k: Using rule (2), construct from the remaining sets S_{k-1}, the sets S_k for phase k by adding to each S_{k-1} an element of $V(S_{k-1})$ to obtain one or more S_k from each S_{k-1}. Eliminate duplicate S_k sets as they occur. As in phase 1, compute λ^{S_k} and $V(S_k)$ for each S_k. Using rule (3), stop if a solution λ^{S_k} is found. As before, remove any set S_k so that $V(S_k) = \phi$ and λ^{S_k} is not the solution. If necessary, proceed to phase k+1.

In the improved Theil and van de Panne the stages are renumbered, so that "stage" corresponds to the number of vertices determining the projection hyperplane—rather than the

number of them omitted from this determination. The "unconstrained projection" is determined at Stage m, which is still done first.

One major improvement takes advantage of the larger memory of the IBM 7094: Inverses of Γ_+ and its required submatrices are all computed and stored before the start of the print-by-print estimation of proportions. Thus, a matrix inverse need no longer be computed for every projection. The λ's at Stage k - 1 can be computed from one of these inverses and the corresponding λ at Stage k, by Theorem 2. These two considerations make for savings in computation time.

The theorems presented above also provide for some simplification in geometric signature analysis (Theorem 6) and the alien object test (Theorems 1, 3, and 6).

The present version of Program MIXMAP was written for the IBM 7094. An explanation of this program is given below. This includes the present computational procedures for geometric signature analysis and the alien object test.

A.3 A DETAILED EXPLANATION OF THE PROGRAM

In Step 1, the Boolean variable FIRST is set = 1B before the statement to read the control data. In Step 2, a long conditional branch starting at "WHENEVER FIRST" reads signatures, makes transformations, and calculates inverses of submatrices of GAMMA. In this loop, FIRST is set equal to OB. Thus, the branch is entered after the first PROCESS input following control input and at no other time.

The entry TRUBL is usually called in POINT modules in case of difficulty. It calls SYSTEM if ITEST = 0 and otherwise calls ERROR, producing an octal dump. In MIXMAP, the function name variable TRUB is set equal to TRUBL., and then TRUB is called in various contingencies. The reason for this roundabout procedure is that if an unexplained bug appears in the program, you can include after the statement "TRUB=TRUBL." a statement "WHENEVER ITEST.G.1, ERROR2." and get a decimal dump to help find what's wrong. A decimal dump could, in that case, be forced by giving NSIG too large a value in the control input. A binary deck for XDUMP is also needed.

The transformations made in the "WHENEVER FIRST" branch of Step 2 are as follows:

The vertices AA(NSIG by ND) and the covariance matrices MM(NSIG by ND by ND) are read by READMX. ND is the number of channels on the data tape, presumed to agree with the number of channels in the signature decks.

The L5 loop forms a new vertex matrix A and cumulates ABAR, the negative of the average vertex, and MBAR, the average covariance matrix, using only the subset of channels chosen. ABAR is then added to each vertex because the new origin of the space is going to be the average

vertex. There are two reasons for making this translation. One is to reduce Δ to 0, as shown in Theorem 6, so that Δ does not have to be calculated and can be left out of the in-plane and out-of-plane distance formulas. The other reason is to ensure an easily invertible GAMMA matrix.

MXSCL. (FSIG, ABAR, ABAR, NX) multiplies the vector ABAR (1) . . . ABAR(NX) by the scalar FSIG, = 1./NSIG, and stores the results in ABAR.

A covariance-removing transformation y = Px is calculated as follows. Let M be the covariance matrix of x. It is a theorem that the covariance matrix of y is PMP^T.

Proof. Cov y = $\mathcal{E}yy^T - \mathcal{E}y\mathcal{E}y^T$

where "$\mathcal{E}$" stands for "expectation of."

$$\mathcal{E}yy^T = \mathcal{E}y\mathcal{E}y^T = \mathcal{E}Pxx^TP^T - \mathcal{E}Px\mathcal{E}x^TP^T$$

P is a constant and comes outside the expectation operator. So

$$\text{cov } y = P(\mathcal{E}xx^T)P^T - P\mathcal{E}x\mathcal{E}x^TP^T = P(\mathcal{E}xx^T - \mathcal{E}x\mathcal{E}x^T)P^T = P\,M\,P^T \quad \text{Q.E.D.}$$

The problem is to find P such that

$$PMP^T = I$$

$$M = P^{-1}P^{T-1}$$

$$M^{-1} = P^TP$$

because the inverse of a product is the product of the inverses in reverse order. Thus P is found by inverting M, making the Cholesky decomposition

$$M^{-1} = P\,P^T$$

and then transposing P.

In the program, MBAR is taken as the best estimate of M, then inverted by GJR with the result stored in MBAR, and then subjected to the Cholesky decomposition subroutine LLT with the result stored in MBAR. Because the result of LLT is a lower triangular matrix, all elements above the diagonal are zeroed and removed. The transpose of MBAR, an upper triangular matrix, is the transformation to apply to the data vector X.

However, to save time in transforming X, the multiplications by zero are omitted. In the double loop early in PROMIX, the point-processing internal function, the last value of y, namely Y(NX) is calculated first. X(NX) is the floating equivalent of the input integer DATUM(NX) and Y is the initial value of Y plus X(NX) times the (NX,NX) value of the transposed MBAR, now called P. Y(NX-1) depends on X(NX), X(NX-1) and two values of P. Doing the loop

backwards this way allows the X's to be calculated as they are used. The initial values of Y are ABAR, the negative of the average vertex, thus effecting a translation to put the origin at the average vertex. The P's are referenced by the linear subscript K to avoid the time to compute a double subscript. Thus P is stored, not in matrix form, but in the unusual linear form required by the backwards loop in PROMIX.

The vertices A and the average vertex ABAR must also undergo the same covariance-removing transformation. Any column vector Z would be transformed PZ, but if Z were a row vector, it would be transformed ZP^T. The vertices are stored in A as row vectors and ABAR can be considered a row vector. Thus A and ABAR are transformed by multiplying them by MBAR, which has been reduced to P^T.

The GAMMA matrix is computed next. GAMMA(I,J) is the inner product of the I-th and J-th rows of A. A column of 1's and a row of 1's are added, and GAMMA(NSIG+1, NSIG+1) is defined as 0. We have found the GAMMA matrix often poorly conditioned, that is, attempted inversion of GAMMA led to trouble. The problem was solved by making, in effect, a change of scale

new Y = old Y/V

where V is the square root of the largest element W of GAMMA in absolute value. The effect on GAMMA is to divide the first NSIG rows and columns by W, with the result that the largest element is now 1. A, ABAR, and P are also divided by V, so that the transformation takes no time at the point-processing stage.

The L20 loop goes through every subset of the vertices, computing and storing the inverse of the corresponding submatrix of GAMMA. A subset is identified by an integer ISUB which, when looked at in binary form, has a 1 corresponding to every retained vertex. A subset of the first, second, and fifth vertex, from right to left, has the number 10011 in binary, which is the same as 19 in decimal. Every subset of five vertices can be represented as an integer from 0 to 31. The L20 loop eliminates subsets of size 0 and 1 because the corresponding inverses are not needed for the projections.

Within the L20 loop, the subset is expanded into the vector ICOL, so that ICOL(I) = 1 if the vertex is present and 0 otherwise. The same vector is also called BCOL so that it can be treated as a Boolean variable as well as an integer. The sum of the vector is the number of vertices in the subset, and is stored in ICOL(0). This number plus 1 is called ICOL1 and is the row length of the corresponding submatrix of GAMMA. It is one more because of the border of 1's. ICOL(NSIG+1) is kept equal to 1 to indicate that this border is always part of the subset. The rows and columns of GAMMA are picked out where ICOL(I) = 1 and stored in C(H+1) . . . C(H+ICOL1*ICOL1). The inverse is then computed using subroutine GJR and stored in the same place. The index H starts at 0 and after every trip through the loop it is

incremented by ICOL*ICOL1. The reason it is not incremented by $ICOL1^2$ is that the last row of the inverse is not used in calculating the projections and therefore does not have to be saved.

To retrieve the inverses that have been computed and stored sequentially in the C vector, a pointer CLOC(ISUB) is saved. Logically, the starting index H should be saved by CLOC. But H - ICOL1 is saved instead, so that the zero-th element of the k-th row of the inverse is the simple expression

C(CLOC(ISUB) + K*ICOL1)

The distance from each vertex to the opposite face can now be obtained from the inverse C of the subset NSUB, the one that includes all the vertices. By Theorem 4, this distance D(K) is $1/\sqrt{C_{KK}}$ for each vertex K. To restore it to the original standard deviation units, it is multiplied by V and printed out as DIST(K). The index of C_{KK} is

CLOC(NSUB) + K*(NSIG+1) + K

which reduces to

CLOC(NSUB) + K*(NSIG+2)

DMIN, the smallest of these distances, is a measure of how poor the estimation of proportions is likely to be. It has been stated that if DMIN were 0, the estimates would be totally ambiguous. If DMIN for a certain K were near 0, then small chance fluctuations in X(K) would cause wild fluctuations in the mixture estimates. The program limit, MINDIS, has a default value of 0.5. Many users, however, may want a more stringent limit, and require, for example, that all vertices be at least one standard deviation away from the opposite face. Then they would set MINDIS = 1 in the control input. Any DMIN less than MINDIS causes the program to stop.

Some Boolean switches are precalculated to speed up the point-processing. BAL means "make the alien object test," BOUT means "compute the out-of-plane distance," BAV means "keep a cumulative sum of the data points" so that mixture estimates can be obtained for the average data point, NOMIX means "do not compute mixture estimates point by point," NOXAL means "compute neither mixture estimates nor alien object tests," and ITEST2 and ITEST3 signal a debugging printout. The average point AVEPT (an integer variable), PRO (the cumulative proportion vector), ALNO (the number of alien objects), and HNO (the number of non-alien objects) are zeroed prior to point-processing.

The point-processing routine is the internal function PROMIX. "WHENEVER NOXAL, TRANSFER TO UPDATE" is used because if there are neither mixture estimates nor alien object tests to calculate, all that remains is to update AVEPT and HNO and return.

The procedure for getting Y has already been described. The floating point equivalent of the integer data vector DATUM is stored in X. Y is computed with the 1/V transformation built into P, and a translation of coordinates to move the origin to the average vertex is effected. The vector G is calculated. G(K) is the inner product of the K-th vertex with Y.

The initial λ values are calculated—that is, the λ's of the projection onto the vertex space. λ_K is just the inner product of the K-th row of C_1, the inverse of GAMMA, with G. If all the λ's are positive, the estimation is complete, and we can transfer to OUT. ICOL, L, SETL and STM are given values first, because these variables are used in the section of program starting with OUT.

In the L110 loop, MINL is computed to be $\min_K \{LAM(K) * D(K)\}$ or 0, whichever is less. D(K), we remember, is $1/\sqrt{C_{KK}}$. Thus by Theorems 3 and 6, $MINL^2$ is the square of the in-plane distance for the alien object test. The out-of-plane distance squared is, by Theorems 1 and 6, the inner product of Y with itself minus the inner product of LAM with G. D2 is the sum of the two squared distances unless no out-of-plane distance is calculated, in which case it is only the square of the in-plane distance.

Theoretically, one would take the square root of D2 and compare it with ALSIG, which is the distance in standard deviations Y has to be from the simplex to be considered alien. $\sqrt{D2}$ is actually a lower bound for the distance. The out-of-plane distance is exact, but the in-plane distance is the largest distance from the point to a face that separates the simplex from the point, and is therefore a lower bound for the true in-plane distance. Thus, if $\sqrt{D2} >$ ALSIG, then the true distance > ALSIG, the point is alien, and the mixture estimates need not be computed. If $\sqrt{D2} <$ ALSIG, then proportions can be calculated, a true distance obtained, and the test reapplied. Thus the alien points, and only they, are rejected. In practice, a precomputed value TEST=$ALSIG^2/W$ is compared with D2 to allow for the 1/V transformation and to avoid taking the square root.

When a point fails the alien object test, the alien object counter is updated, DATUM(NSIG+2) is set = 1 (indicating the point is alien), DATUM(NSIG+1) is set = D2, and DATUM(1) . . . DATUM(NSIG) are given the conventional value 777_8.

When a point passes the alien object test and the option of only calculating mixture estimates from the average point has been selected, then there is nothing left to do but transfer to UPDATE, where the average point and the good point counter are updated and the function returns.

The two-page TVLOOP carries out the Thiel and van de Panne procedure with a few modifications. The loop passes in stages from NSIG to 2. At each stage ST, it is assumed that the λ's, namely LAM(J), have been computed and stored sequentially, and the action at stage ST is

to compute the λ's for the next lower stage, and while doing that to test signs to see if all λ's are positive. If so, the optimality test is performed, and if this test succeeds, then the estimation procedure is over.

λ values from only two stages need to be kept in storage: those at stage ST that are being referenced, and those at the next lower stage that are being computed. The LAM array where these λ's are stored is in two halves. The λ's computed for stage NSIG before entering the TVLOOP are stored starting at LAM(O) while those computed at stage NSIG for stage NSIG-1 are stored starting at LAM(350). After stage NSIG has been completed, and all λ's computed for stage NSIG-1 (and have been tested and found wanting) then the stage NSIG λ's are no longer needed and at stage NSIG-1, the λ's for stage NSIG-2 are computed and stored starting at LAM(O).

At stage ST the sets of λ's that have been computed are gone through using the index J. J starts at J1, which is initially 0 and is subtracted from 350 after every stage. Thus J1 alternates between 0 and 350. Preceding each set of λ's in the LAM vector is a word giving the subset to which the λ's refer. This word is given the integer name SET(J) and made equivalent to the floating point array LAM. Thus, at the outset SET(O) = NSUB and the initial λ's are stored in LAM(1) . . . LAM(NSIG).

The JLOOP going through the sets of λ's is almost as big as the TVLOOP. The increment is by ST+1 to allow for the defining subset SET(J) as well as the ST λ's stored compactly, without gaps. The subset is expanded into ICOL(1) . . . ICOL(NSIG), where ICOL(N) = 1 if the N-th vertex is present in the subset, and 0 otherwise.

Within the JLOOP (and almost as big) is the KLOOP, going through the λ_K's for K = 1 . . . ST. Actually the loop goes for N = 1 . . . NSIG and only executes when ICOL(N) = 1, at which point K is incremented. In this way, both the index K of the λ's, which are stored compactly, and the index N of the G vector, which is not, are available for use in the KLOOP.

We are looking within the KLOOP at a value of N for which ICOL(N) = 1 (i.e., BCOL(N) = 1B). SETL is the original subset SETJ with column N zeroed. This is accomplished by bitwise intersection of SETJ with COL(N), where COL is an integer vector 1, 2, 4, 8, 16, . . . , which is in binary form, 00001, 00010, 00100, 01000, etc. The sign of λ_K is tested. If it is negative, then the Thiel and van de Panne procedure designates SETL as a possible subset on which the nearest point to Y might be found. Therefore λ's are computed for SETL, using Theorem 2 and the inverse of the submatrix of GAMMA located at CLOC(SETJ).* The λ's are stored by the index L in the half of the LAM array not used by J. This is accomplished

*Theorem 2 states that new λ_i = old λ_i - C_{ik} * (λ_K/C_{KK}), where it is the K-th vertex that has been deleted.

by starting L at 350-J1 outside the JLOOP. SET(L) is given the value SETL, and the new λ's are stored in LAM(L=1) . . . LAM(L+ST-1). Then L is incremented by ST.

If the sign of λ_K is not negative, then it appears that subset SETL need no longer be considered; if any subset of SETJ contains the nearest point to Y, then that subset will be found in one of the SETL's corresponding to a negative λ. Thus as far as succeeding subsets of λ at stage ST are concerned, SETL is done and need not be looked at again. Either it was a possible subset and λ's were calculated for it, or it was a subset thrown out because of corresponding to a positive λ. Thus before the test of the sign of λ_K, it is recorded that SETL is bypassed by setting DONE(SETL)=1B. The DONE array starts out all zero. In every subsequent pass through the KLOOP, when DONE(SETL) is found to be 1B, then that pass through the KLOOP is bypassed.

The new λ's are tested for sign as they are computed. If all are positive then the optimality test is performed. This consists of going through the empty columns I of SETL (loop L120), forming SETI by adding column I to SETL, locating the inverse matrix C corresponding to SETI, and then computing λ_I directly as the inner product of the NK-th row of the inverse with G. NK is found by counting the 1's of SETI from right to left and stopping after the 1 in column I. It is needed because the inverse is stored compactly. Whenever $\lambda_I > 0$, the test has failed and a transfer to FAIL is made. If loop L120 is completed without failure, then the last λ's computed define the nearest point on the simplex to Y and a transfer to OUT is made.

When a subset SETL fails the optimality test, then the point on SETL nearest to Y has been found and it is not optimal. That means that neither SETL nor any of its subsets can be optimal, so we might as well record them all as being done. The section of code from FAIL to the END OF CONDITIONAL following L125 does this job. The method is to form all subsets of SETL consisting of 1 vertex each, and then combine them in all possible ways, recording DONE=1B for each combination. (The best way to understand this section of code is to work it through with an example subset.)

The TVLOOP finishes with ST=2 when λ's for subsets of size 1 are calculated. According to theory, some subset must have had positive λ's and passed the optimality test, because there is always some point on the simplex that is nearest to Y. The program allows for the possibility that because of roundoff or a program bug, no optimal subset can be found. A message is then printed giving the line and point number so that the point can be examined later. The point is counted as alien. If ITEST ≧ 2, the run ends there with a dump. (Occurrences of this message would be of interest to the authors.)

To minimize the possibility that no subset will be found optimal, the tests of sign are made not with zero but very small numbers: Z = 0.000001 and ZM = -0.0000005. A set of λ's is considered all positive if they are all >ZM rather than all >0. To fail the optimality test, λ_i has to be >Z rather than >0.

The section of code from OUT up to STEP(5) describes the action taken after the optimal subset has been found. The most recently calculated set of λ's, which are stored compactly, are spaced out and stored in the vector PROP with zeroes stored in the spots corresponding to the missing vertices. The true distance squared, $\sum(y_i - \lambda_i A_i)^2$, is calculated and compared with TEST to see for sure whether the point is alien. (The alien object test represents a screening operation. Every point failing the test is alien, but some points passing the test will also be alien.) If so, the alien object counter ALNO is updated and DATUM(NSIG+2) is set = 1.

If the point passes this second alien object test, the good point counter HNO is updated and DATUM(NSIG+2) is set = 0. If a mixture estimate will be made from the average data point, then the cumulative sum of data points AVEPT(1) . . . . . AVEPT(NX) is updated. The cumulative sum of mixture estimates PRO(1) . . . . . PRO(NSIG) is updated. The squared distance is converted to standard deviation units by multiplying by W and then scaled into the range 0 . . . 511 by multiplying by 5. The last multiplication is done first in Step 2 resulting in a factor W5. The number of output channels NX is set = NSIG+2. The point-processing routine PROMIX has now completed its work so it returns.

In Step 5, the branch of the MIXMAP module called after the points of the requested rectangle have been processed, the total number of points, the number HNO of good points, and ALNO of alien points, and the time to process the points are printed. If mixture estimates have been calculated, then the sum vector PRO of the mixture estimates is divided by the number of good points and the result stored in the vector PROP. Internal function PRTMIX is called to print the estimates.

If mixture estimates based on an average of the data points have been requested, then the sum vector AVEPT of the data points is divided by the number of good points, converted to an integer and stored in DATUM. We need now only call PROMIX to get, in the vector PROP, the mixture estimates based on the average data point. We must first be sure to turn off the alien object test, the average point cumulation, and the no-mixture-estimate bypass. The distance from the average point to the simplex in standard deviation units is computed by taking the square root of DIST*W and is then printed.

Appendix B
PROOFS OF THEOREMS RELATING TO AVERAGING PROCEDURES

The purpose of this appendix is to derive results for use in Section 4 on averaging procedures with the standard estimator, and/or in Section 5.1 on averaging procedures with the simplified proportion estimator.

We will assume the positive integers m and n are fixed with $m \leq n + 1$; here, m represents the number of pure signatures and n represents the number of spectral channels. S denotes the fundamental m-1 simplex, i.e., the set of all points in the positive octant of Euclidean m space $\mathcal{E}^m$, with components which sum to one. Equivalently, S is the convex hull of the unit vectors.

Material i of the m materials is assumed to have a signature with mean A_i and covariance matrix M_i. A will denote the matrix which has A_i for its i-th column. Let λ be a column vector representing a mixture of materials, i.e., the j-th component λ^j of λ is the proportion of material j, $1 \leq j \leq m$. It is clear that λ is in S. Such a vector is termed a proportion vector. The ERIM model for signatures of mixtures states that the mean A_λ and covariance matrices M_λ of mixture of materials with proportion vector λ are given by:

$$A_\lambda = A\lambda$$

$$M_\lambda = \sum_{j=1}^{m} M_j \lambda^j$$

Let y denote a signal obtained from a pixel containing materials with associated true proportion vector λ_o (unknown). The standard estimator $\hat{\lambda}_o$ of λ_o is defined as follows. The set of points

$$A\lambda \quad \lambda \text{ in } S$$

form a simplex S_A which is called the signature simplex. Let Z be the point in S_A which is closest (in the Euclidean metric) to y. Then Z can be represented as

$$Z = A\hat{\lambda}_o$$

where $\hat{\lambda}_o$ is a proportion vector. $\hat{\lambda}_o$ is taken to be an estimate of the true proportion vector λ_o. $\hat{\lambda}_o$ is unique if the signature simplex is nondegenerate, i.e., if S_A has positive (m-1)-dimensional volume. We will always assume this condition is satisfied. If the pure signatures are Gaussian and their covariance matrices are equal, then after an appropriate preprocessing transformation of the pure signatures and data points, $\hat{\lambda}_o$ is the maximum likelihood estimate of λ_o.

Let us consider the following situation. Suppose there are N data points y_i, $1 \leq i \leq N$, each representing a pixel with true proportion vector λ_i. The region represented by the aggregate of the N pixels would then have true proportion $\bar{\lambda}$ given by

$$\bar{\lambda} = \frac{1}{N}\sum_{i=1}^{N} \lambda_i$$

Here we are assuming pixels of equal size. One may estimate a proportion vector $\hat{\lambda}_i$ for each i-th pixel with signal y_i. Then the average of these estimates $\hat{\lambda}_i$ given by

$$\bar{\hat{\lambda}} = \frac{1}{N}\sum_{1}^{N} \hat{\lambda}_i$$

may be taken as an estimate of $\bar{\lambda}$. The mean square error of the estimate is then given by

$$E\,||\bar{\hat{\lambda}} - \bar{\lambda}||^2$$

where E is the expected value operator and $||\cdot||$ represents the Euclidean norm. For each i, $1 \leq i \leq N$, let

$$b_i = E(\hat{\lambda}_i) - \lambda_i$$

Then b_i is the bias of the estimate $\hat{\lambda}_i$. Let

$$\bar{b} = \frac{1}{N}\sum_{1}^{N} b_i$$

Then $\bar{b}$ is the average bias of the $\hat{\lambda}_i$. Assuming the y_i are statistically independent, the following result may be derived.

Theorem 1: $$E\,||\bar{\hat{\lambda}} - \bar{\lambda}||^2 = \frac{1}{N^2}\sum_{1}^{N} E\,||\hat{\lambda}_i - E(\hat{\lambda}_i)||^2 + ||\bar{b}||^2$$

Proof:

$$E\,||\bar{\hat{\lambda}} - \bar{\lambda}||^2 = E\,||\bar{\hat{\lambda}} - E(\bar{\hat{\lambda}}) + E(\bar{\hat{\lambda}}) - \bar{\lambda}||^2 = E\,||\bar{\hat{\lambda}} - E(\bar{\hat{\lambda}})||^2 + ||E(\bar{\hat{\lambda}}) - \bar{\lambda}||^2$$

$$= \frac{1}{N^2} E\left|\left|\sum (\hat{\lambda}_i - E(\hat{\lambda}_i)\right|\right|^2 + ||\bar{b}||^2 \qquad \text{Q.E.D.}$$

An alternative method for estimating $\bar{\lambda}$ is the following. Let

$$\bar{y} = \frac{1}{N}\sum_{1}^{N} y_i$$

denote the average of the signals. Treat $\bar{y}$ as a signal from a pixel and estimate a proportion vector $\hat{\bar{\lambda}}$ and take $\hat{\bar{\lambda}}$ as the estimate of $\bar{\lambda}$. Then the mean-square error of $\hat{\bar{\lambda}}$ given by

$$E||\hat{\lambda} - \bar{\lambda}||^2 \leqq \frac{\beta\tau}{N}$$

satisfies (still under the assumption that the y_i are statistically independent),

Theorem 2: $E||\hat{\lambda} - \bar{\lambda}||^2 \leqq \frac{\beta\tau}{N}$

where

$$\tau = \max_i \text{trace } M_i$$

$$1 \leqq i \leqq m$$

and β depends upon A but not N.

The proof of Theorem 2 depends upon three lemmas.

Lemma 1: Let y be a data point and let $\hat{\lambda}$ be the standard estimate of the true proportion vector λ corresponding to y. Then

$$||A\hat{\lambda} - A\lambda|| \leqq ||y - A\lambda||$$

Let L denote the linear hull of the fundamental simplex S; i.e., L is the set of vectors with components summing to one. Clearly, L contains S, the fundamental simplex. The linear hull L_A of the signature simplex S_A is defined to be the set of points in signal space of the form

$A\eta$ for η in L

Lemma 2: Let L be the linear hull of S with A nondegenerate. Then a β (depending on A) such that

$$||\eta - \xi||^2 \leqq \beta||A\eta - A\xi||^2$$

for all η and ξ in L. (β may be taken as the reciprocal of the smallest eigenvalue of B^TB where the $(n + 1) \times (m)$ matrix B is obtained from A by the addition of a row of 1's. The fact that B^TB is nonsingular is a consequence of the fact that the signature simplex is nondegenerate.)

Lemma 3: Let M_i, $1 \leqq i \leqq m$ be positive definite symmetric matrices. Let y be a vector-valued random variable with mean zero and covariance matrix M_λ given by

$$M_\lambda = \sum_{j=1}^{m} M_j \lambda^j$$

where λ is in S and λ^j denotes its j-th component. Then

$$E||y||^2 \leqq \tau$$

where

$$\tau = \max_i \text{trace } M_i$$

$$1 \leq i \leq m$$

Proof of Lemma 1: Consider the triangle with vertices y, $A\hat{\lambda}$, and $A\lambda$. Let α denote the angle at vertex $A\hat{\lambda}$. Since $A\hat{\lambda}$ and $A\lambda$ are both in S_A, every point on the line segment between them is in S_A. If $\alpha < \frac{\pi}{2}$ radians, there would be a point $A\lambda_o$ on this segment closer to y than $A\hat{\lambda}$. This would contradict the choice of $\hat{\lambda}$. Therefore $\alpha \geq \frac{\pi}{2}$ radians, and consequently the length of the largest side of the triangle is $||y - A\lambda||$. Since $||A\hat{\lambda} - A\lambda||$ is the length of another side of this same triangle, the lemma follows.

Proof of Lemma 2: Let B be the matrix obtained from A by adding a column of 1's. Since A is nondegenerate,

B^TB is nonsingular

Let η and ξ be in L and let $\omega = \eta - \xi$. Then the sum of the components of ω is zero and it follows that

$$A^TA\omega = B^TB\omega$$

Then

$$||A\eta - A\xi||^2 = ||A\omega||^2 = \langle A^TA\omega, \omega\rangle = \langle B^TB\omega, \omega\rangle \geq \alpha||\omega||^2$$

where α is the smallest eigenvalue of B^TB. This last equality is a well known property of positive definite symmetric matrices. Thus

$$||\eta - \xi||^2 = ||\omega||^2 \leq \frac{1}{\alpha}||A\eta - A\xi||^2 \qquad \text{Q.E.D.}$$

Proof of Lemma 3:

$$E||y||^2 = \text{trace } M_\lambda = \text{trace} \sum_1^m \lambda^j M_j = \sum_1^m \lambda^j (\text{trace } M_j) \leq \tau \sum_1^m \lambda^j = \tau \qquad \text{Q.E.D.}$$

Proof of Theorem 2: There are positive numbers β and τ such that

$$E||\hat{\lambda} - \bar{\lambda}||^2 \leq \beta E||A\hat{\lambda} - A\bar{\lambda}||^2 \qquad \text{(by Lemma 2)}$$

$$\leq \beta E||\bar{y} - A\bar{\lambda}||^2 \qquad \text{(by Lemma 1)}$$

(Equation Continued)

$$= \beta E\left|\left|\frac{1}{N}\sum_{i=1}^{N}(y_i - A\lambda_i)\right|\right|^2 = \frac{\beta}{N^2}\sum_{1}^{N} E||y_i - A\lambda_i||^2$$

$$\leq \frac{\beta\tau}{N} \qquad \text{(by Lemma 3)} \qquad \text{Q.E.D.}$$

We now turn our attention to theorems concerning averaging procedures using the simplified proportion estimator $\tilde{\lambda}$. As before, let L denote the linear hull of the fundamental simplex S and let L_A denote the linear hull of the signature simplex S_A. Recall that for a point η in L, $\eta^{\oplus}$ denotes the vector obtained by setting all negative components of η equal to zero. Let $\rho = \rho(\eta)$ denote the sum of the positive components of η. Note that $\rho \geq 1$ because the sum of all the components of η is 1. Now let $\tilde{\eta} = \frac{1}{\rho}\eta^{\oplus}$. Note that $\tilde{\eta}$ is a proportion vector.

Let Py denote the orthogonal projection of the signal y onto L_A, i.e., Py is the unique point in L_A closest to y. Py has a representation

$$Py = A\eta$$

for some η in L. This representation is unique for nondegenerate S_A. If λ_o is the true proportion vector of the pixel represented by the signal y, then the simplified estimator $\tilde{\lambda}_o$ of λ_o is defined to be

$$\tilde{\lambda}_o = \tilde{\eta}$$

Some properties of $\tilde{\lambda}$ with respect to averaging procedures will be examined. These properties are similar to the corresponding averaging properties of the estimator $\hat{\lambda}$. As before, let y_i, $1 \leq i \leq N$, denote a signal from a pixel with true proportion vector λ_i. The region represented by the aggregate of the N pixels would then have true proportion vector $\bar{\lambda}$ given by

$$\bar{\lambda} = \frac{1}{N}\sum_{i=1}^{N}\lambda_i$$

Here we are assuming pixels of equal size. If we estimate $\bar{\lambda}$ by averaging the estimates $\tilde{\lambda}_i$ from the individual pixels, we obtain the estimate

$$\bar{\tilde{\lambda}}_i = \frac{1}{N}\sum_{i=1}^{N}\tilde{\lambda}_i$$

Let

$$\tilde{b}_i = E(\tilde{\lambda}_i) - \lambda_i$$

and let $\tilde{\bar{b}}$ be defined by

$$\tilde{\bar{b}} = \frac{1}{N}\sum_{1}^{N}\tilde{b}_i$$

Then $\tilde{\bar{b}}$ is the average bias of the estimators $\tilde{\lambda}_i$, and the mean-square error of the estimator $\tilde{\bar{\lambda}}$ is given by

$$E||\tilde{\bar{\lambda}} - \bar{\lambda}||^2$$

Assuming the y_i to be statistically independent, we have Theorem 3:

$$E||\tilde{\bar{\lambda}} - \bar{\lambda}||^2 = \frac{1}{N^2}\sum_{1}^{N}||(\tilde{\lambda}_i - E(\tilde{\lambda}_i)||^2 + ||\tilde{\bar{b}}||^2$$

Proof:

$$E||\tilde{\bar{\lambda}} - \bar{\lambda}||^2 = E||\tilde{\bar{\lambda}} - E(\tilde{\bar{\lambda}}) + E(\tilde{\bar{\lambda}}) - \bar{\lambda}||^2 = E||\tilde{\bar{\lambda}} - E(\tilde{\bar{\lambda}})||^2 + ||E(\tilde{\bar{\lambda}}) - \bar{\lambda}||^2$$

$$= E\left|\left|\frac{1}{N}\sum(\tilde{\lambda}_i - E(\lambda_i))\right|\right|^2 + ||\tilde{\bar{b}}||^2 = \frac{1}{N^2}\sum E||\tilde{\lambda}_i - E(\lambda_i)||^2 + ||\tilde{\bar{b}}||^2 \qquad \text{Q.E.D.}$$

Now let

$$\bar{y} = \frac{1}{N}\Sigma y_i$$

and let $\tilde{\bar{\lambda}}$ denote the simplified proportion estimate for the signal $\bar{y}$ and consider $\tilde{\bar{\lambda}}$ as an estimate of $\bar{\lambda}$. Then

Theorem 4: $E||\tilde{\bar{\lambda}} - \bar{\lambda}|| \leq \frac{m\beta\tau}{N}$

where m is the number of pure signatures

$$\tau = \max_i \text{trace } M_i$$

$$1 \leq i \leq m$$

and β depends upon A but not N.

The proof of this theorem requires the following additional lemma:

Lemma 4: For all η in L and λ in S

$$||\tilde{\eta} - \lambda||^2 \leq m||\eta - \lambda||^2$$

Proof of Lemma 4: Let $\eta \in L$ and $\lambda \in S$. If $\eta \in S$, then $\tilde{\eta} = \epsilon$ and the result follows. Therefore, we assume that $\eta \notin S$. Therefore η has some negative components. Let ν denote the number of negative components of η. Then since η also has positive components, it follows that

$$1 \leq \nu \leq m - 1 \tag{B-1}$$

Let η^{Θ} denote the vector obtained by setting the non-negative components of η equal to zero. Then

$$\eta = \eta^{\oplus} + \eta^{\Theta} \tag{B-2}$$

$$\eta^{\oplus} \text{ is orthogonal to } \eta^{\Theta} \tag{B-3}$$

Let λ_1 denote the vector obtained from λ by setting the j-th, $1 \leq j \leq m$, component of λ equal to zero when the corresponding j-th component of η is negative. Let λ_2 denote the vector obtained from λ by setting the j-th, $1 \leq j \leq m$, component of λ equal to zero when the corresponding j-th component of η is non-negative. Then

$$\lambda = \lambda_1 + \lambda_2 \tag{B-4}$$

$$\lambda_1 \text{ is orthogonal to } \lambda_2 \tag{B-5}$$

$$||\lambda_1||^2 \leq 1 \tag{B-6}$$

$$||\lambda_2||^2 \leq 1 \tag{B-7}$$

$$\lambda_1 \text{ is orthogonal to } \eta^{\Theta} \tag{B-8}$$

$$\lambda_2 \text{ is orthogonal to } \eta^{\oplus} \tag{B-9}$$

Let ρ be the sum of the positive components of η and let $\rho = 1 + \epsilon$. Then

$$\epsilon > 0 \tag{B-10}$$

Now,

$$||\tilde{\eta} - \lambda||^2 = \left|\left|\frac{\eta^{\oplus}}{\rho} - \lambda\right|\right|^2 \qquad \text{(by def. of } \tilde{\eta})$$

$$= \left|\left|\frac{\eta^{\oplus}}{\rho} - \lambda_1 - \lambda_2\right|\right|^2 \qquad \text{(by Eq. B-4)}$$

$$= \left|\left| \frac{\eta^{\Theta}}{\rho} - \lambda_1 \right|\right|^2 + ||\lambda_2||^2 \qquad \text{(by Eqs. B-5 and B-9)}$$

$$= \frac{1}{(1+\epsilon)^2} \left|\left| \eta^{\Theta} - \lambda_1 - \epsilon\lambda_1 \right|\right|^2 + ||\lambda_2||^2$$

$$< \left|\left| \eta^{\Theta} - \lambda_1 - \epsilon\lambda_1 \right|\right|^2 + ||\lambda_2||^2$$

$$\leq \left[\left|\left| \eta^{\Theta} - \lambda_1 \right|\right| + \epsilon||\lambda_1|| \right]^2 + ||\lambda_2||^2$$

$$= \left|\left| \eta^{\Theta} - \lambda_1 \right|\right|^2 + 2\epsilon||\lambda_1|| \cdot \left|\left| \eta^{\Theta} - \lambda_1 \right|\right| + \epsilon^2||\lambda_1||^2 + ||\lambda_2||^2$$

$$\leq \left|\left| \eta^{\Theta} - \lambda_1 \right|\right|^2 + 2\epsilon \left|\left| \eta^{\Theta} - \lambda_1 \right|\right| + \epsilon^2 + ||\lambda_2||^2 \qquad \text{(by Eq. B-6)}$$

Therefore,

$$||\tilde{\eta} - \lambda||^2 \leq \left|\left| \eta^{\Theta} - \lambda_1 \right|\right|^2 + 2\epsilon \left|\left| \eta^{\Theta} - \lambda_1 \right|\right| + \epsilon^2 + ||\lambda_2||^2 \qquad \text{(B-11)}$$

Now consider $||\eta^{\Theta}||^2$. The sum of the components of η^{Θ} is $-\epsilon$. Therefore, the minimum value of $||\eta^{\Theta}||^2$ occurs when η^{Θ} has each of its ν nonzero components equal to $\frac{-\epsilon}{\nu}$. Thus,

$$||\eta^{\Theta}||^2 \geq \frac{\epsilon^2}{\nu}$$

and from Eq. (B-1), it follows that

$$||\eta^{\Theta}||^2 \geq \frac{\epsilon^2}{m-1} \qquad \text{(B-12)}$$

Now,

$$||\eta - \lambda||^2 = \left|\left| \eta^{\Theta} + \eta^{\Theta} - \lambda_1 - \lambda_2 \right|\right|^2 \qquad \text{(by Eqs. B-2 and B-4)}$$

$$= \left|\left| \eta^{\Theta} - \lambda_1 \right|\right|^2 + \left|\left| \eta^{\Theta} - \lambda_2 \right|\right|^2 \qquad \text{(by Eqs. B-3, B-5, B-8, B-9)}$$

$$= \left|\left| \eta^{\Theta} - \lambda_1 \right|\right|^2 + ||\eta^{\Theta}||^2 - 2\langle \eta^{\Theta}, \lambda_2 \rangle + ||\lambda_2||^2$$

$$\geq \left|\left| \eta^{\Theta} - \lambda_1 \right|\right|^2 + \left|\left| \eta^{\Theta} \right|\right|^2 + \left|\left| \lambda_2 \right|\right|^2$$

$$\geq \left|\left| \eta^{\Theta} - \lambda_1 \right|\right|^2 + \frac{\epsilon^2}{m-1} + ||\lambda_2||^2 \qquad \text{(by Eq. B-12)}$$

Therefore,

$$||\eta - \lambda||^2 \geqq \left|\left|\eta^{\Theta} - \lambda_1\right|\right|^2 + \frac{\epsilon^2}{m-1} + ||\lambda_2||^2 \tag{B-13}$$

Then,

$$m||\eta - \lambda||^2 - ||\tilde{\eta} - \lambda||^2 \geqq (m-1)\left|\left|\eta^{\Theta} - \lambda_1\right|\right|^2 - 2\epsilon\left|\left|\eta^{\Theta} - \lambda_1\right|\right| + \frac{1}{m-1}\epsilon^2$$

$$+ (m-1)||\lambda_2||^2 \qquad \text{(by Eqs. B-11 and B-13)}$$

$$\geqq \left[(m-1)^{1/2}\left|\left|\eta^{\Theta} - \lambda_1\right|\right| - \left(\frac{1}{m-1}\right)^{1/2}\epsilon\right]^2$$

$$\geqq 0 \qquad \text{Q.E.D.}$$

R. Kauth has pointed out that m is not the best possible bound. He has outlined a proof that

$$||\tilde{\eta} - \lambda||^2 \leqq C_m||\eta - \lambda||^2$$

where

$$C_m = \begin{cases} \dfrac{(1+m)^2}{4m} & m \text{ odd} \\ \dfrac{m+2}{4} & m \text{ even} \end{cases}$$

and that C_m is the best possible bound.

Proof of Theorem 4: Let P denote the orthogonal projection of signal space onto L_A, the linear hull of the signature simplex S_A. Let L denote the linear hull of all proportion vectors. Then

$$Py_i = A\eta_i \quad \text{for some unique } \eta_i \text{ in L}, \quad 1 \leqq i \leqq N$$

Set

$$\bar{\eta} = \frac{1}{N}\sum_{1}^{N}\eta_i$$

Then $P\bar{y} = A\bar{\eta}$ and there are positive-number β and τ such that

$$E||\tilde{\bar{\lambda}} - \bar{\lambda}||^2 = E||\tilde{\bar{\eta}} - \bar{\lambda}||^2$$

$$\leqq mE||\bar{\eta} - \bar{\lambda}||^2 \qquad \text{(by Lemma 2)}$$

(Equation Continued)

$$\leq m\beta E\,||A\bar{\eta} - A\bar{\lambda}||^2 \qquad \text{(by Lemma 4)}$$

$$= m\beta E\,||P\bar{y} - A\bar{\lambda}||^2$$

$$\leq m\beta E(||P\bar{y} - A\bar{\lambda}||^2 + ||y - P\bar{y}||^2)$$

$$= m\beta E\,||y - A\bar{\lambda}||^2$$

The last equality follows from the fact that $y - P\bar{y}$ is orthogonal to L_A, which contains $P\bar{y}$ and $A\bar{\lambda}$. Continuing, we have

$$E||\tilde{\bar{\lambda}} - \bar{\lambda}||^2 \leq m\beta E\,||y - A\bar{\lambda}||^2$$

$$= m\beta E\left|\left|\frac{1}{N}\sum_1^N (y_i - A\lambda_i)\right|\right|^2$$

$$= \frac{m\beta E}{N^2} E\sum_1^N ||y_i - A\lambda_i||^2$$

$$\leq \frac{m\beta\tau}{N} \qquad \text{Q.E.D.} \qquad \text{(by Lemma 3)}$$

Appendix C
DESCRIPTION OF MIXTURES DATA SIMULATION PROGRAM

To facilitate testing of the ERIM proportion-estimation algorithm, a computer program was written to generate simulated multispectral data of the type that would come from a mixture of materials. This was necessary to insure that completely reliable "ground truth" was available for each data point, so that meaningful mean-square errors of estimation would be calculated over large areas. Written for the IBM 7090, this program is described in detail below.

The mixtures data generated by Program SIM are based on the ERIM mixtures model, and may be either pure mixtures of known materials or may contain a proportion of alien material. Any signature set may be used to generate the data, including signature tapes written on the CDC 1604. The actual distribution of alien materials and the proportion vectors are governed by parameters set by the user. The program is written as an External Function in MAD on the IBM 7094.

C.1 THE MIXTURES DATA MODEL

No matter whether the pixel contains a "pure" material or a mixture, it is assumed that the signal s from any given pixel is a Gaussian random variable with mean vector A and covariance matrix M. The ERIM mixtures model relates the statistics of the pure signatures to the statistics of the mixture according to the formulae

$$A = \sum_{i=1}^{m} \lambda^i A_i = \text{mixture mean}$$

$$M = \sum_{i=1}^{m} \lambda^i M_i = \text{mixture covariance}$$

where λ^i is the proportion of the i-th "pure" material in the pixel, A_i its signature mean, and M_i its signature covariance matrix. This can be extended to the case where there are two classes of materials, each with its own proportions λ^i over the class. Let there be t materials in the first class and u materials in the second class. Let the pixel be completely covered by the two classes and let ξ be the proportion of the second class in the pixel. Then the model implies that

$$A = (1 - \xi)\sum_{i=1}^{t} \lambda^i A_i + \xi \sum_{i=1}^{u} \lambda^{i+t} A_{i+t}$$

$$M = (1 - \xi)\sum_{i=1}^{t} \lambda^i M_i + \xi \sum_{i=1}^{u} \lambda^{i+t} M_{i+t}$$

The two classes might be taken to be "user" and "alien" materials, where "user" materials are known materials for which suitable signatures are available. Then ξ is the proportion of alien material in the pixel. Note that ξ and the λ^i are themselves random variables over a collection of pixels.

For each pixel, we desire to "randomly" choose A and M such that they specify an appropriate mixtures distribution. Before this can be done, it is necessary to choose random values of ξ and the λ^i. Given the number of materials in a pixel, the λ^1 are to be uniformly distributed. The number of "user" signatures or "alien" signatures actually employed to generate the data is also a random variable and is taken to have a certain distribution over the set (1, 2, . . . , t). The distribution function F(x) of ξ used in the program is

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ \alpha + (1 - \alpha - \beta)\dfrac{1 - e^{-\gamma x}}{1 - e^{-\gamma}} & \text{for } 0 \le x \le 1 \\ 1 & \text{for } x \ge 1 \end{cases}$$

where α is the probability that the pixel contains "user" material only, β is the probability that it contains "alien" material only, and γ is another parameter.

For each pixel, the program must determine randomly how many "user" materials and how many "alien" materials are in it, and this requires some prior statistical information. It would suffice to specify the quantities

$$\rho_i^u \text{ and } \rho_j^A$$

which are the probabilities that the pixel contains exactly i "user" materials and j "alien" materials. In an agricultural application, these quantities are related through the geometry of the field configuration and can be computed from the parameters τ_u and τ_A. Both of these represent an average ratio of field length to pixel edge, but one is used for the ρ_i^u and the other for the ρ_j^A. The following formulas were derived:

$$\rho_1(\tau) = (1 - \tau)^2$$

$$\rho_2(\tau) = 2\tau - 2.5\tau^2$$

$$\rho_3(\tau) = \tau^2$$

$$\rho_4(\tau) = 0.5\tau^2$$

assuming that there can be no more than four fields in a pixel. These formulas can be used for either ρ_A of ρ_u. In the existing program we have added (for "user" materials)

$$\rho_5(\tau) = 0.25\tau^2$$

$$\rho_6(\tau) = 0.25\tau^2$$

since, for a very small field size, it might be possible to have as many as six fields overlap in one pixel. Program SIM allows a maximum of six user materials and three alien materials per pixel. For both alien and user materials, the $\rho_i(\tau)$ are normalized to sum to 1. Once these ρ_i are determined, the program can randomly choose the numbers t and u of materials in each pixel and then "randomly" choose the proportions λ^i.

Given the extended mixtures model

$$A = (1 - \xi)\sum_{i=1}^{t}\lambda^i A_i + \xi\sum_{i=1}^{u}\lambda^{i+t}A_{i+t}$$

$$M = (1 - \xi)\sum_{i=1}^{t}\lambda^i M_i + \xi\sum_{i=1}^{u}\lambda^{i+t}M_{i+t}$$

the A_i and M_i, and the random quantities ξ and λ, the mixture signature mean A and covariance M can be computed. A sample from the Gaussian distribution with mean A and covariance M is taken as the signal vector for the pixel.

C.2 A DESCRIPTION OF EXTERNAL FUNCTION SIM

All of the programming required to generate the data is done in SIM, which is written as an External Function. Writing the data out on tape in standard ERIM multispectral format is accomplished by OUT 94, a special output routine. Both are called by a small enabling program which first reads certain data (see below) and instructs the operator to mount a signature tape from which the simulated mixtures data are to be generated. The first call to SIM (from the enabling program) computes certain variables and sets up variables required for linkage to OUT94. Succeeding calls are made by OUT94 to Internal Function LINE (part of SIM) in order to generate the data vector for each point. The data vector, which is partly the generated signal vector and partly a vector of "ground truth" proportions, is packed by a special subroutine. When a complete packed line of data has been generated, it is passed back to OUT94 to be written out on tape.

In order to interface with Assembly Line, External Function SIM contains the Assembly Line Common Block of Q-variables. In the first part of the program certain of these Q-vari-

ables are set, and certain SIM variables are set to zero or their default values. The variables SCAL, MODE, IPRINT, SUNIT, IPRIN2 are then read via a READ AND PRINT DATA statement, and a data title is read in. The signatures for the "user" and "alien" materials are read in via READMX, with SCAL = 51.2 if CDC-1604 signatures are used. Finally, the ρ_i are computed—the first six for "user" signatures and the last three for the "alien" signatures. These are normalized to sum to 1. The arguments for the random number generators (RND for uniform and RNP for normal distribution) are initialized to zero and SIM then returns to the enabling program.

The remaining portion of SIM consists entirely of Internal Function LINE, which generates a complete packed line of multispectral data. This is a program entry to SIM and is called by OUT94 each time a line is to be generated. If the line number requested is too large, LINE returns to OUT94 where an end-of-file is written on the tape. Otherwise it starts by setting up the arguments required for PAK so that the packed output line can be stored in lower Erasable. The random number generator is "warmed up" by exercising it 50 times. LINE then enters a loop which generates the individual data points for the line requested by OUT 94. This loop, which ends at ENDL, does the following:

(1) Determines proportion ξ of alien material in the pixel.
(2) Selects number of "user" materials that will occur in the pixel.
(3) Selects number of "alien" materials that will occur in the pixel.
(4) Randomly chooses the subset of materials for (2).
(5) Randomly chooses the subset of materials for (3).
(6) Generates a partial proportion vector, λ_u, corresponding to "user" materials.
(7) Generates a partial proportion vector, λ_A, corresponding to "alien" materials.
(8) From (1), (6), and (7), computes the total proportion vector, $\lambda = (1 - \xi)\lambda_u + \xi\lambda)_A$.
(9) Using λ and the signature statistics A_i and M_i, computes the mixture signature mean A and covariance M.
(10) Chooses a sample X from the normal distribution determined by A and M. (This requires a Cholesky decomposition and matrix multiplication.)
(11) Forms combined vector (x, λ) and "packs" it in ERIM output format using PAK.
(12) Goes to next point; if last point is done, calls final entry to PAK.
(13) Returns to OUT94 and writes packed line on tape.

The number of components for λ is always 9, although some may be zero. The number of data channels may be as large as 15, so the total vector of data plus ground truth proportions may have as many as 24 components on the output. When OUT94 has written the last line, it returns to the enabling program, which then terminates.

How to Use SIM

The program deck for generating simulated data must include the following:

- A source deck for the enabling program (described below)
- Binary decks for SIM, PAKH, and READMX
- Data (described below)

The enabling program must be of the form

*MAIN PROGRAM TO ENABLE SIM
N'R
P'N Q(225)
"Dimension," "Erasable" statements
F'T__, __, __. . .
READ AND PRINT DATA
Mount signature tape on SUNIT
SIM.(NSIG, NCHAN, NSS, TAU1, TAU2, NUSER, NALIEN, ALFA, BETA, GAMMA)
OUT94. (F, L, S, UNIT, BIN1, BIN2)
E'M

and the arguments to SIM and OUT94 are read in via the READ AND PRINT DATA statement. SIM itself has certain "input" variables which need not be read in since they are set to default values. These are

SCAL (default = 1.0)
MODE (default = 1)
IPRINT (default = OB
IPRIN2 (default = OB

and they can be changed via the READ AND PRINT DATA statement of SIM.

A list of input variable definitions follows:
NSIG = number of signatures used (NALIEN and NUSER)
NCHAN = number of data channels to be used
NSS = number of points per line generated by SIM
TAU1 = value of τ for computing ρ_i for "user" signatures
TAU2 = value of τ for computing ρ_i for "alien" signatures
NUSER = number of user signatures to be read
NALIEN = number of alien signatures to be read
ALFA, BETA, GAMMA = parameters used to calculate ξ (defined previously)
F = first line number
L = last line number

S = line number interval

SUNIT = tape drive for signature (input) tape

UNIT = tape drive for output tape

BIN1, BIN2, bin location of first and last output tapes

SCAL = scale factor for reading signatures via READMX (should be 51.2 for 1604 tape)

MODE = MODE argument of READMX

IPRINT, IPRINT2 = Boolean switches for providing debugging output from SIM

The data deck consists of a $DATA card plus input corresponding to two READ AND PRINT DATA statements. The first is from the enabling program and specifies values for the variables

NSIG, NCHAN, NSS, F, L, S, UNIT, BIN1, BIN2, TAU1, TAU2, NUSER, NALIEN, ALFA, BETA, GAMMA

The second is from SIM itself and may specify any of the following:

SCAL, MODE, IPRINT, IPRIN2, SUNIT

which will be set to default values if not specified in the output. After these, the input deck must include a card specifying a title for the simulated data. The title may occupy up to 72 columns. An output tape with a write ring inserted into the back of it must be mounted on the tape drive whose number is specified by the variable UNIT. Data is generated by SIM at a rate of about 0.5 sec per point.

C.3 GENERAL REMARKS

This program can provide data for testing any of several recognition or estimation techniques, even though it does not provide for a geometric configuration "fields." It might be desirable to modify SIM so that more than one data point can be sampled from each mixture distribution.

REFERENCES

1. Horwitz, H. M., R. F. Nalepka, P. D. Hyde, and J. P. Morgenstern, 1971, Estimating the Proportions of Objects Within a Single Resolution Element of a Multispectral Scanner, Seventh International Symposium on Remote Sensing of Environment, Report No. 10259-1-X, May 1971, Willow Run Laboratories of the Institute of Science and Technology, The University of Michigan, Ann Arbor.

2. Nalepka, R. F., H. M. Horwitz, and P. D. Hyde, 1972, Estimating Proportions of Objects From Multispectral Data, Report No. 31650-73-T, Willow Run Laboratories of the Institute of Science and Technology, The University of Michigan, Ann Arbor.

3. Nalepka, R. F., and P. D. Hyde, 1973, Estimating Crop Acreage From Space-Simulated Multispectral Scanner Data, Report No. 31650-148-T, Environmental Research Institute of Michigan, Ann Arbor.

4. Malila, W. A., and R. F. Nalepka, 1973, Atmospheric Effects in ERTS-1 Data, and Advanced Information Extraction Techniques, Symposium On Significant Results Obtained From the Earth Resources Technology Satellite-1, Vol. 1, Goddard Space Flight Center, Greenbelt, Md.

5. Thomson, F. J., 1973, Crop Species Recogniton and Mensuration in the Sacramento Valley, Symposium On Significant Results Obtained From the Earth Resources Technology Satellite-1, Vol. 1, Goddard Space Flight Center, Greenbelt, Md.

6. Salvato, Jr., P., 1973, Iterative Techniques to Estimate Signature Vectors For Mixture Processing of Multispectral Data, Conference on Machine Processing of Remotely Sensed Data, Purdue University, Lafayette, Indiana.

7. Detchmendy, D. M., and W. H. Pace, 1972, A Model for Spectral Signature Variability for Mixtures, Earth Resources Observation and Information Analysis Systems Conference, Tullahoma, Tennessee.

8. Kunzi, H. P., W. Krella, and W. Oettli, 1966, Nonlinear Programming, Blaisdell Publishing Co., Boston.

9. Malila, W. A., R. H. Hieber, D. P. Rice, and J. E. Sarno, 1973, Wheat Classification Exercise, Using June 11, 1973, ERTS MSS Data for Fayette County, Illinois (For CITARS Task), Report No. 190100-21-R, Environmental Research Institute of Michigan, Ann Arbor.

DONOT PRINT

DISTRIBUTION LIST

NASA/Johnson Space Center
Earth Observations Division
Houston, Texas 77058

ATTN: Dr. A.Potter/TF3 (4)
ATTN: Mr. Robert MacDonald (1)
ATTN: Mr. B. Baker/TF3 (8)
ATTN: Mr. R. Dean Bretton/TF53
Chief Earth Resources Research Data Facility (8)
ATTN: Mr. A. H. Watkins/HA (1)

NASA/Johnson Space Center
Facility & Laboratory Support Branch
Houston, Texas 77058

ATTN: Mr. D. Riley BB631/B4 (1)

NASA/Johnson Space Center
Computation & Flight Support
Houston, Texas 77058

ATTN: Mr. Eugene Davis/FA (1)

NASA Headquarters
Washington, D.C. 20546

ATTN: Mr. C. W. Mathews (1)

U.S. Department of Agriculture
Agricultural Research Service
Washington, D.C. 20242

ATTN: Dr. Robert Miller (1)

U.S. Department of Agriculture
Soil & Water Conservation Research Division
P.O. Box 267
Weslaco, Texas 78596

ATTN: Dr. Craig Wiegand (1)

U.S. Department of Interior
Geological Survey
Washington, D.C. 20244

ATTN: Dr. James R. Anderson (1)

Director, Remote Sensing Institute
South Dakota State University
Agriculture Engineering Building
Brookings, South Dakota 57006

ATTN: Mr. Victor I. Myers (1)

U.S. Department of Interior
Fish & Wildlife Service
Bureau of Sport Fisheries & Wildlife
Northern Prairie Wildlife Research Center
Jamestown, North Dakota 58401

ATTN: Mr. Harvey K. Nelson (1)

U.S. Department of Agriculture
Forest Service
240 W. Prospect Street
Fort Collins, Colorado 80521

ATTN: Dr. Richard Driscoll (1)

U.S. Department of Interior
Geological Survey
Water Resources Division
500 Zack Street
Tampa, Florida 33602

ATTN: Mr. A. E. Coker (1)

U.S. Department of Interior
Director, EROS Program
Washington, D.C. 20244

ATTN: Mr. J. M. Denoyer (1)

Earth Resources Laboratory, GS
Mississippi Test Facility
Bay St. Louis, Mississippi 39520

ATTN: Mr. R. O. Piland, Director (1)
ATTN: Mr. Sid Whitley (1)

U.S. Department of Interior
Geological Survey
GSA Building, Room 5213
Washington, D.C. 20242

ATTN: Mr. W. A. Fischer (1)

NASA Wallops
Wallops Station, Virginia 23337

ATTN: Mr. James Bettle (1)

Purdue University
Purdue Industrial Research Park
1200 Potter
West Lafayette, Indiana 47906

ATTN: Dr. David Landgrebe (1)
ATTN: Dr. Philip Swain (1)
ATTN: Mr. Terry Phillips (1)

U.S. Department of Interior
EROS Office
Washington, D.C. 20242

ATTN: Dr. Raymond W. Fary (1)

U.S. Department of Interior
Geological Survey
801 19th Street, N.W.
Washington, D.C. 20242

ATTN: Mr. Charles Withington (1)

U.S. Department of Interior
Geological Survey
801 19th Street, N.W.
Washington, D.C. 20242

ATTN: Mr. M. Deutsch (1)

U.S. Geological Survey
801 19th Street, N.W., Room 1030
Washington, D.C. 20242

ATTN: Dr. Jules D. Friedman (1)

U.S. Department of Interior
Geological Survey
Federal Center
Denver, Colorado 80225

ATTN: Dr. Harry W. Smedes (1)

U.S. Department of Interior
Geological Survey
Water Resources Division
901 S. Miami Ave.
Miami, Florida 33130

ATTN: Mr. Aaron L. Higer (1)

University of California
School of Forestry
Berkeley, California 94720

ATTN: Dr. Robert Colwell (1)

School of Agriculture
Range Management
Oregon State University
Corvallis, Oregon 97331

ATTN: Dr. Charles E. Poulton (1)

U.S. Department of Interior
EROS Office
Washington, D.C. 20242

ATTN: Mr. William Hemphill (1)

Chief of Technical Support
Western Environmental Research Laboratories
Environmental Protection Agency
P.O. Box 15027
Las Vegas, Nevada 89114

ATTN: Mr. Leslie Dunn (1)

NASA/Langley Research
Mail Stop 470
Hampton, Virginia 23365
ATTN: Mr. William Howle (1)

U.S. Geological Survey
Branch of Regional Geophysics
Denver Federal Center, Building 25
Denver, Colorado 80225
ATTN: Mr. Kenneth Watson (1)

NAVOCEANO, Code 7001
Naval Research Laboratory
Washington, D.C. 20390
ATTN: Mr. J. W. Sherman, III (1)

U.S. Department of Agriculture
Administrator
Agricultural Stabilization and Conservation Service
Washington, D.C.
ATTN: Mr. Kenneth Frick (1)

Pacific Southwest Forest & Range Experiment Station
U.S. Forest Service
P.O. Box 245
Berkeley, California 94701
ATTN: Mr. R. C. Heller (1)

Pacific Southwest Forest & Range Experiment Station
U.S. Forest Service
P.O. Box 245
Berkeley, California 94701
ATTN: Dr. P. Weber (1)

NASA/Johnson Space Center
Mission Planning & Analysis Division
Houston, Texas 77058
ATTN: Mr. H. G. De Vezin/FM8 (1)

University of Texas at Dallas
Box 688
Richardson, Texas 75080
ATTN: Dr. Patrick L. Odell (1)

Department of Mathematics
University of Houston
Houston, Texas 77004
ATTN: Dr. Henry Decell (1)

Institute for Computer Services and Applications
Rice University
Houston, Texas 77001
ATTN: Dr. M. Stuart Lynn (1)

U.S. National Park Service
Western Regional Office
450 Golden Gate Avenue
San Francisco, California 94102
ATTN: Mr. M. Kolipinski (1)

U.S. Department of Agriculture
Statistical Reporting Service
Washington, D.C. 20250
ATTN: D. H. VonSteen/R. Allen (2)

U.S. Department of Agriculture
Statistical Reporting Service
Washington, D.C. 20250
ATTN: Mr. H. L. Trelogan, Administrator (1)

Department of Watershed Sciences
Colorado State University
Fort Collins, Colorado 80521
ATTN: Dr. James A. Smith (1)

Lockheed Electronics Co.
16811 El Camino Real
Houston, Texas 77058
ATTN: Mr. R. Tokerud (1)

TRW System Group
Space Park Drive
Houston, Texas 77058
ATTN: Dr. David Detchmendy (1)

IBM Corporation
1322 Space Park Drive
Houston, Texas 77058
ATTN: Dr. D. Ingram (1)

S&D—DIR
Marshall Space Flight Center
Huntsville, Alabama 35812
ATTN: Mr. Cecil Messer (1)

Code 168-427
Jet Propulsion Laboratory
4800 Oak Grove Drive
Pasadena, California 91103
ATTN: Mr. Fred Billingsley (1)

NASA/Johnson Space Center
Technical Library Branch
Houston, Texas 77058
ATTN: Ms. Retha Shirkey/JM6 (4)

NASA Headquarters
Washington, D.C. 20546
ATTN: Mr. W. Stoney/ER (1)
ATTN: Mr. Leonard Jaffe/ER (1)
ATTN: Mr. M. Molloy/ERR (1)
ATTN: Mr. G. Thorley/ERR (1)

Ames Research Center
National Aeronautics and Space Administration
Moffett Field, California 94035
ATTN: Dr. I. Poppoff (1)

Goddard Space Flight Center
National Aeronautics and Space Administration
Greenbelt, Maryland 20771
ATTN: Mr. W. Nordberg, 620 (1)
ATTN: Mr. W. Alford, 563 (1)

Lewis Research Center
National Aeronautics and Space Administration
21000 Brookpark Road
Cleveland, Ohio 44135
ATTN: Dr. Herman Mark (1)

John F. Kennedy Space Center
National Aeronautics and Space Administration
Kennedy Space Center, Florida 32899
ATTN: Mr. S. Claybourne/FP (1)

NASA/Langley
Mail Stop 214
Hampton, Virginia 23665
ATTN: Mr. James L. Raper (1)